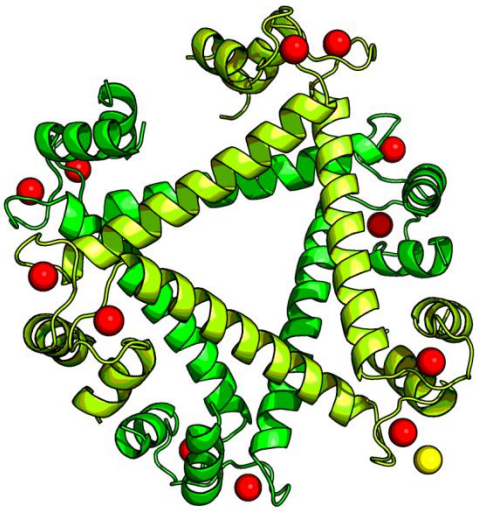
PREDICTION, DESIGN AND ANALYSIS OF CANONICAL EF-HAND LOOP AND QUALITATIVE ESTIMATION OF Ca^{2+} BINDING AFFINITY



Thesis submitted to Jawaharlal Nehru University
for the award of degree of

DOCTOR OF PHILOSOPHY

By
MOHIT MAZUMDER

School of Life Sciences
Jawaharlal Nehru University
New Delhi – 110067
INDIA

SCHOOL OF LIFE SCIENCES

Jawaharlal Nehru University

New Delhi – 110067

INDIA

CERTIFICATE

This is to certify that the research work embodied in this thesis entitled as **"Prediction, Design and analysis of canonical EF-hand loop and qualitative estimation of Ca^{2+} binding affinity"** submitted for the award of degree of **Doctor of Philosophy**, has been carried out by **Mr. Mohit Mazumder** under the guidance and supervision of Prof. Samudrala Gourinath, Professor, Structural Biology Laboratory at the School of Life Sciences, Jawaharlal Nehru University, New Delhi, India.

The work is original and has not been submitted so far, in part or full for the award of any degree or diploma of any other university.

Mohit Mazumder

(Ph.D. Candidate)

Prof. Samudrala Gourinath

(Supervisor)

for PKatti 26/7/17

Prof. S. K. Goswami

(Dean)

कार्यकारी डीन / Acting Dean
जीवन विज्ञान संकाय / School of Life Sciences
जवाहरलाल नेहरू विश्वविद्यालय
Jawaharlal Nehru University
नई दिल्ली/New Delhi-110067

SCHOOL OF LIFE SCIENCES
Jawaharlal Nehru University
New Delhi – 110067
INDIA

CERTIFICATE OF ORIGINALITY

The research work embodied in this thesis entitled as **"Prediction, Design and analysis of canonical EF-hand loop and qualitative estimation of Ca^{2+} binding affinity"** has been carried out by me at the School of Life Sciences, Jawaharlal Nehru University, New Delhi, India. The thesis has been subjected to plagiarism check by **'Tunitin'** software. The work submitted for the consideration of award of Ph.D. is original.

Mohit Mazumder

SCHOOL OF LIFE SCIENCES

Jawaharlal Nehru University

New Delhi – 110067

INDIA

Supervisor's Certificate for Exclusion of Self-Published work

The content of the Chapter 2 entitled **"Prediction and Analysis of Canonical EF-Hand Loop and Qualitative Estimation of Ca^{2+} Binding Affinity"** have been published as

"Prediction and Analysis of Canonical EF Hand Loop and Qualitative Estimation of Ca^{2+} Binding Affinity. Mohit Mazumder · Narendra Padhan · Alok Bhattacharya · Samudrala Gourinath* · PLoS ONE Apr 2014

The published work has been included in the thesis and has not been submitted for any degree to any University/ Institute.

Mohit Mazumder

(Student)

Prof. Samudrala Gourinath

(Supervisor)

Dedicated to...

Maa Bapi and Bonu

Acknowledgement

I will start with "The PhD life", I joined school of life sciences in 2009 and as you can see it has been a long time. I joined lab-430 as a JRF to work under Dr. Samudrala Gourinath as my supervisor. To have completed my research project under his guidance has been an invaluable experience. It is really difficult to put all of these experiences into words but for me it has been a blessing and I consider myself very lucky to have sir as my PhD supervisor. He has provided the direction, the facility and opportunities for me to follow my research project successfully with lot of freedom ☺ to work and collaborate as well. His scientific intuitions, high standards, genuine involvement and contributions to the research being carried out by each and every member of our lab, has been inspiring. I have learnt a lot from him both professionally and personally. It is indeed rare for a supervisor to maintain a friendly atmosphere and a professional aura in lab; sir you have done both and that too at their best. The ease with which we could talk to you about everything from professional to personal problems is exceptional. Your forgiveness for our silly mistakes in lab has been my inspiration which had taught me to be patient with my juniors. All the efforts that you had put in teaching me all the chemistry and towards the understand of structure biology is highly acknowledged. I think I can write a whole chapter describing "How cool sir is": D. Dear Sir, I truly cannot thank you enough for the amazing PhD experience. It has been an enthralling academic journey; we will have more than 20 peer reviewed research papers together. Thanks, are due to Dr. Neelima Alam too for her loving and caring nature with all of us. We have shared amazing memories while going

to all of those lab trips with ma'am, sir and cute little Mansha. I specifically remember the last one which was our trip to Jaisalmer, the trip was like a movie with all credits to the Bus driver due thanks to Nitesh for saving us all that while. :p.

My heartfelt thanks to Prof. Alok Bhattacharya, School of Life Sciences, for helping me with the calcium binding project his views and points of understanding of scientific problems were amazing. I would like to thank him for giving me the opportunity to teach as well.

I would like to thank all my research collaborators as I have learned so much by working with them on so many different proteins and problems. I would like to thank Prof Prof. Rohini Muthuswami, Prof S.S. Komath, Prof Alok Mondal, Prof. Supriyo Chakorborty for their inputs and valuable suggestions during my pre-and post PhD duration.

I am grateful to the previous Deans of SLS, Prof. P.K. Yadava, Prof. R. Madhubala, Prof. Mallick, Prof. B. C. Tripathy and the current Dean Prof. S.K.Goswami for providing the necessary infrastructure and facility at the School for carrying out my research. I am also thankful to all the faculty members at SLS for extending their support and scientific inputs during my work presentations and otherwise too. It is a pleasure to acknowledge Dr. Surya Prakash Sharma of former SLS library, Suneeta ma'am and all the SLS office staff and administration office. I would like to thank Ved Prakash for being there as a friend , Karan and krishna for laboratory assistance.

I gratefully acknowledge Indo shahtri institute for the Shastri Research Student award, NVIDIA for student professor award and fujitshu for C1020 GPU card award.

I would like to thank Sanjeev and devbrat for helping me with the calcium binding wet lab experiments. They have also inspired me to learn the wet lab myself :p. I would like to thank Deepak, Nitesh and Sanjeev for being there all the time. I would like to acknowledge Shivesh, Isha, Sudhir, Arif, Faisal, Nitesh, Rohit, Sudhakar, Gunjan, Suneeta, Preety Satya, Poonam, Dhaka, Priya, Pragyan, Preeti Umrao, Tamanna, Vijayan, Shadab, Amir for being the best people.

I would like to thank JNU and SLS TT team for all that I have achieved it is because of you guys. I would also like to acknowledge the SLS cricket team for letting me represent our school. It's impossible to forget the jhadi cricket experience all thought my PhD. We have been at our childhood best during all these times playing table tennis, cricket, Badminton and carom board.

Friends are angels who lift us to our feet when our own wings have trouble remembering how to fly. I would have to list like 100 people here to start with instead I would like to thank all of you. Our life's shapes up like the way we really want I believe in that. These people have changed my life by sharing all life experiences together.

I would especially like to thank Ankita Dutta for being there through all the difficult times. You have been inspiration for me to do well in every aspect of my life. It would have been a shaky thesis compilation if you were not around. Thank you for your never-ending support.

With the genuine love and companionship of my friends outside the lab, have made this PhD experience truly remarkable. that I will never forget. These are the precious memories that I ll never forget. The JNU experience, the Jawarharlal Nehru university, I just fell in love with this place is has always been my second home. It is an era for me

living all those years in this beautiful place where I see peacocks after every rain. I would thank like to thank everyone in this University.

No amount of expressing my gratitude and love for the support of my family could ever equal their natural unconditional love and understanding for me. My family is my rock of support and strength. Thank you, Maami and Bapi for never understanding my reasons for not being there on various occasions; but still supporting me in my endeavor, thanks to you and Bonu for doubling my fellowship as I remember how hard it is get fellowship money in JNU. The research scholar pain as they all say. Thanks to my little sis for her love and care and support.

Thankyou. Last, but the most important, hard to believe in, the unseen force behind the existence of the universe ...

"Fire is His head, the sun and moon His eyes, space His ears, the Vedas His speech, the wind His breath, the universe His heart. From His feet the Earth has originated. Verily, He is the inner self of all beings."

~ Anonymous, The Upanishads

Abstract.

Calcium signalling is an integral part of all the biological systems. Calcium homeostasis in cells is regulated by different Ca^{2+} binding proteins which have differential binding affinities for Ca^{2+} ion. The binding affinities of small ions to the proteins can be experimentally determined, howsoever, each method has certain limitations. Due to this it's not always possible to experimentally determine Ca^{2+} binding properties of EF-hand containing Calcium binding proteins (CaBPs). In this scenario it is imperative to predict this property from primary sequences using computational approaches. The focus of this study was to annotate correctly canonical EF-hand motif and further classify these on the basis of their Ca^{2+} binding affinities using Support Vector Machine kernel classifiers. The primary sequences of canonical EF-hand loop were taken from PDB to develop a precise and accurate classifier to classify Ca^{2+} binding loops with non-Ca^{2+} binding regions of EF-hand proteins. Using binary and amino acid composition features we achieved 100% accuracy through 5-fold cross validation. Next, we proposed a novel *ab initio* method to predict the calcium binding affinity, where training datasets were generated on the basis of evolutionary information (PSSM scores). The best performing classifier with concatenated features of accessibility and hydrophilicity showed an accuracy of 87% on experimental test data set. Furthermore, we achieved 100% accuracy on an independent dataset obtained from recently published affinity observations. To investigate further, we performed a proteome wide prediction for *E. histolytica* and classified known EF-hand proteins, and found many probable Ca^{2+} binding sites. We compared our results with published pattern search method on *E. histolytica* proteome and demonstrated our method to be more specific and accurate for predicting potential canonical Ca^{2+} binding loops. Utilizing the developed method, we applied two new scoring schemes to refine the prediction further and designed a EF-hand loop unique to the protein database which was capable of binding calcium with high affinities by mutating residues on the basis of machine learned classifier. The unique sequence was incorporated in the *Entamoeba histolytica* Calcium binding protein1 (*Eh*CaBP1) EF-hand loop 2 using site directed mutagenesis. The sixty-six amino acid residues long protein containing modified second EF-hand loop and the *Eh*CaBP1-Wt with low affinity loops were studied for calcium binding properties using ITC calorimetry. The binding energy indicated at ~535-fold increase in the association constant (K_a) of the designed protein compared to the *Eh*CaBP1-Wt.

Furthermore, we used X-ray crystallography to understand the changes at the atomic level leading to changes in the functional behavior of EF-hand motif in terms of calcium binding. Surprisingly, we found out the high-resolution structure that diffracted at 1.9Å, showed a shrinkage in Ca^{2+} binding coordination sphere resulting in strong coordination yielding high affinity for calcium and forming a hexamer due to the structural changes caused by the designed high affinity calcium sequence.

CAL-EF-AFi can therefore be used to accurately and precisely scan proteomes of organisms for potential Ca^{2+}-binding sites of EF-hand proteins and estimate their probable relative binding affinities. We integrated two scoring techniques in the earlier developed method, to design and validated our findings using biochemical, structural and computational techniques. The coordinates obtained after the X-ray diffraction of Nt*Eh*CaBP1 EF2 mutant have been deposited in RCSB protein databank (PDB Code 5XOP).

The results predicted by the theoretical model were validated by experimental studies. Variation from the EF-hand consensus sequence can be used to predict qualitative Ca^{2+}-binding features. However, this may not be sufficient to understand the overall characteristics of CaBPs. The EF-hand motifs assemble to form a lobe (one partner affects the binding affinity of the other) and the Mg^{2+} affinities are not considered in this work due to limitation of experimental data available till date. Future plans include developing an even better algorithm with more information available from the literature. We hope that an increase in the availability of experimental data will help generate a more robust model.

The mutational analysis was carried out by the CalEFAffi2 program. The source of the program is available at http://202.41.10.46/calb/resources.html. The webserver is free accessible for everyone and is available at http://202.41.10.46/calb/. The program is optimal for scanning large protein databases for calcium binding site identification and estimation of binding affinity. The PSM_{LogL} and SVM_{MAR} scores are provided to assist binding affinity modulation for the scientific community working on numerous proteins still to be annotated. The webserver requires only the protein sequence for the prediction without prior knowledge of structural or biochemical information.

Abbreviations and symbols.

A	Alanine
A_{600}	Absorbance at 600nm
Å	Angstrom
AAC	Amino acid composition
AC	Accessibility
ACC	Accuracy
ANP	Atrial Natriuretic Peptide
α	Alpha
AtCBL2	*Arabidopsis thaliana* Calcineurin B-like protein
AUC	Area under the ROC Curve
bp	Base pair
β	Beta
BLASTp	Protein Basic Local Alignment Search Tool
BLOSUM62	BLOcks SUbstitution Matrix62
$^{45}Ca^{2+}$	Calcium-45 radioisotope
CaBPs	Calcium binding proteins
CaM	Calmodulin
CD-HIT	Cluster Database with High Identity with Tolerance
χ^2	Chi-square
$CaCl_2$	Calcium chloride
C-terminus	Carboxy-terminus
CUDA	Compute Unified Device Architecture
D	Aspartate
DCCM	Dynamic Cross Residue Correlation
DNA	Deoxyribo Nucleic Acid

D3	Positive dataset with high PSSM scores
D4	Negative dataset
D5	Test dataset
D6	Independent dataset of binding affinities
D7	Evaluation dataset
DSSP	Definition of Secondary Structure of Protein
E	Glutamate
EDTA	Ethylene diamine tetraacetate
EGTA	Ethylene glycol-bis (β-aminoethyl ether)-N, N,N',N'-tetraacetic acid
EhCaBPs	EF-hand containing calcium binding proteins
E.coli	*Escherichia coli*
E. histolytica	*Entamoeba histolytica*
F	Phenylalanine
FPLC	Fast Protein Liquid Chromatography
FN	False negative
FP	False positive
g	gram
γ	Gamma
H	Hydrogen
HC	Hydrophilicity
HEPES	(4-(2-hydroxyethyl)-1-piperazineethanesulphonic acid)
HYC	Hydrophobicity
HMM	Hidden Markov Model
IPTG	Isopropyl-β-D-thiogalactopyranoside
ITC	Isothermal titration calorimetry
K	Lysine
K_a	Association constant
K_d	Dissociation constant

Kb	Kilobase
K Da	Kilo-dalton
K	Kelvin
λ	Wavelength
MCC	Matthews's correlation co*EF*ficient
MD	Molecular Dynamics
ml	Milliliter
mM	Millimolar
MPD	2-Methyl-2,4-pentanediol
N	Asparagine
NCBI	National Center for Biotechnology Information
n	Stoichiometry
nm	nanometer
N-terminus	Amino terminus
ng	Nanogram
OD	Optical density
PAGE	Poly acrylamide gel electrophoresis
PBS	Phosphate Buffer Saline
PCR	Polymerase Chain Reaction
PDB	Protein Data Bank
PKC	Protein Kinase C
PMCA	Plasma membrane Ca^{2+}-ATPase
PROSITE	Proteins family and domains database maintained at EMBL
PSSM	Position-Specific Scoring Matrix
Ψ-hand	Pseudo *EF*-loop
Q	Glutamine
RBF	Radial basis function
RMSD	Root-Mean Square Deviation

rpm	Rotations per minute
ROC	Receiver Operating Curve
RT	Room Temperature
SDS	Sodium Dodecyl Sulphate
SDM	Site-Directed Mutagenesis
SN	Sensitivity
SP	Specificity
SPR	Surface Plasmon Resonance
SVM	Support Vector Machine
TP	True positive
TN	True negative
3D	Three dimensional
M	Molar
µl	Micro liter
µM	Micro molar
µm	Micro meter
˚C	Degree Celsius
ΔH	Enthalpy change
ΔG	Gibbs' free energy
ΔS	Entropy change

Table of Contents

Chapter 1.

Introduction of calcium binding proteins and review of various EF-hand containing calcium binding proteins and their binding affinity

1.1 Abstract

Calcium an alkaline earth metal is naturally found in its elementary state. The divalent ion (Ca^{2+}) plays important roles in almost all biological systems. Interestingly, it interacts with numerous proteins resulting in the initiation and regulation of large number of physiological processes. The Ca^{2+}-ion binds specifically in the selective sites in the proteins which are conserved across different proteins present in the cell. The specific role of calcium depends on the Ca^{2+}-ion concentration in the intra- and extracellular compartments of the cell. Inside the cell, it plays role in the metabolic regulation, muscle contraction, cell motility, nerve transmission, cell division and growth, secretion and membrane permeability. The selectivity of Ca^{2+} binding proteins (CaBPs) over other physiologically relevant metals is important for their function. The high level of intracellular Mg^{2+}, for example, as compared to Ca^{2+} imposes the necessity of discrimination against Mg^{2+} for Ca^{2+}-binding proteins operating inside the cell. The abundance of calcium binding residues is from mostly turn/loop like structures. One of the expected reasons is due to the flexible nature of loop/turn and also the ability to supply a large number of bulky amino acids from a short stretch of protein sequence. The natural environment does put certain limitations to the range of Ca^{2+} affinities which are compatible with the particular biological role of each protein. Slightest variations in the sequences of calcium binding proteins can cause improper binding affinities in CaBPs can lead to disease such as osteoporosis, Alzheimer's and heart diseases. In this chapter, we have thoroughly reviewed the EF-hand containing CaBPs and the characteristic of calcium binding affinities.

1.2 The diverse role of calcium in cellular processes

Calcium is an alkaline earth metal which is naturally found in its elementary state. The divalent ion (Ca^{2+}) plays important roles in almost all biological systems. Interestingly, the small divalent ion interacts with numerous proteins resulting in the initiation and regulation of large number of physiological processes [1, 2]. These interactions involve major conformational changes in the proteins invoking diverse functions [3]. The proteins involved in calcium-mediated interactions have different metal binding properties. The importance of calcium is evident by its role in the bio mineralization in bones, shells and teeth of all the higher animals [4].

Ca^{2+} ion binds specifically at conserved sites in the proteins which are present in diverse number of proteins inside the cell. It is an intracellular secondary messenger and its concentration changes swiftly during cell stimuli. Its specific role depends on the Ca^{2+}-concentration in the intra- and extracellular compartments of the cell. Inside the cell, it plays role in the metabolic regulation, muscle contraction, cell motility, nerve transmission, cell division and growth, secretion and membrane permeability [5]. It plays a critical role in the blood-clotting processes. Ca^{2+} ion, vitamin K and fibrinogen are involved in the clotting cascade. The Ca^{2+} binding enzyme, epidermal growth factor (EGF) complex, bind to the phospholipid membrane. Ca^{2+} is indispensable for the clotting process [6]. The cells have extracellular calcium reserve and an intracellular reserve which is required for calcium signalling. The concentration of calcium varies in each cellular compartment as shown in figure 1.1. The spatio-temporal changes in the calcium ion concentration in different cellular compartments affect the regulation of cellular signaling [7]. In plants such as *Arabidopsis*, the calcium binding proteins (CaBPs) are involved in Ca^{2+} mediated signalling which assist in plant responses and enable in development of stress-resistance[8].

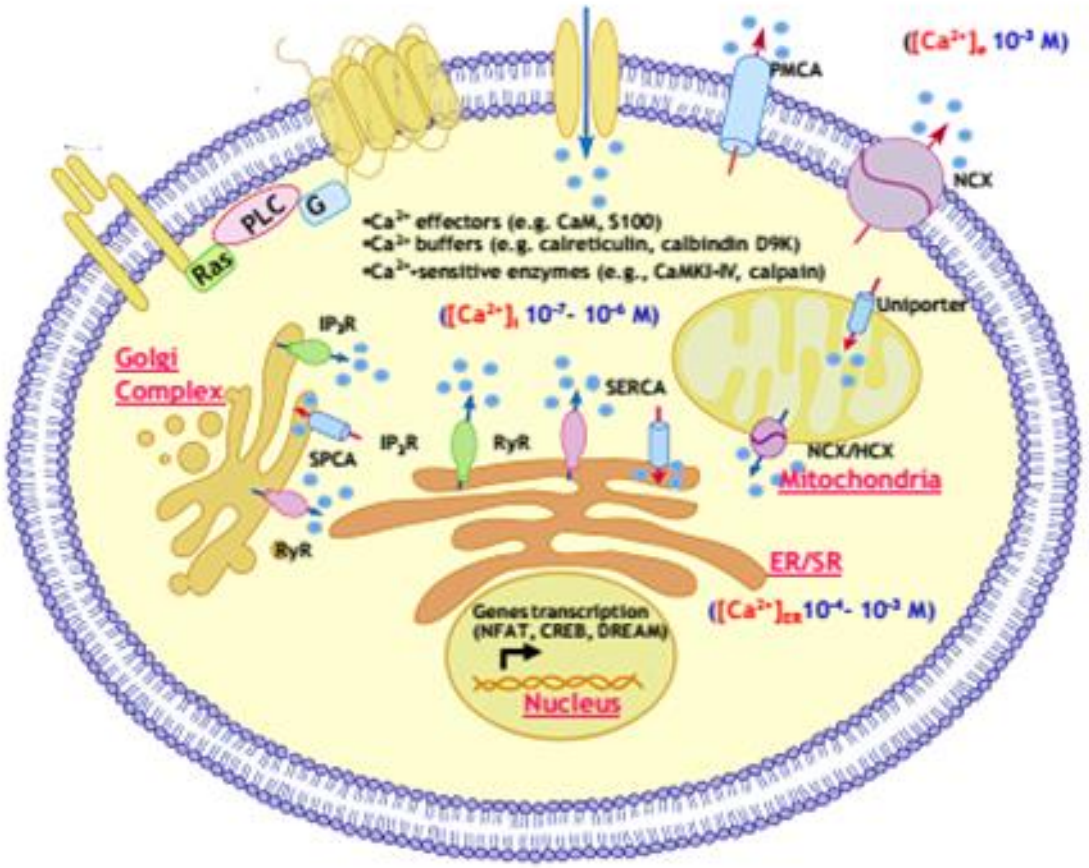


Figure1.1. The figure shows the changes in the calcium ion (Ca^{2+}) concentration upon the extracellular stimulus via various receptors that transfer the extracellular calcium across the plasma membrane and regulates the signalling [7]. (Adapted from Jones *et al* 2006)

1.3. The role of the divalent ion "Calcium" in signalling

Ca^{2+}-ion mediated signalling is regulated by cellular influx of extracellular Ca^{2+} -ions and its *EF*flux from the internal reserves. The channels such as voltage-operated, receptor-operated and many more are involved in the transfer of calcium. Ca^{2+}- ion is also present in the internal stores in the endoplasmic reticulum (ER), lysosomes and Golgi complex which act as reservoirs through IP_3R (inositol-1,4,5-triphosphate receptors) and ryanodine receptor. The mechanism to maintain calcium at the resting state is achieved by plasma membrane via plasma membrane Ca^{2+}-ATPase (PMCA) and sodium-calcium (Na^{+}-Ca^{2+}) exchanger[5, 9].

Alternatively, it is achieved by pumping Ca^{2+} back to internal storage mediated by Ca^{2+}-dependent ATPase. Signalling is further transmitted when Ca^{2+} binds to the calcium signalling molecules such as CaM (Calmodulin), calcium buffer, effector and calcium dependent enzymes. The fate of the signalling cascade is dependent on the interacting protein partners which can have a long-term effect by controlling the activity of various transcription factors [10].

1.4. Classification of Calcium Binding Proteins

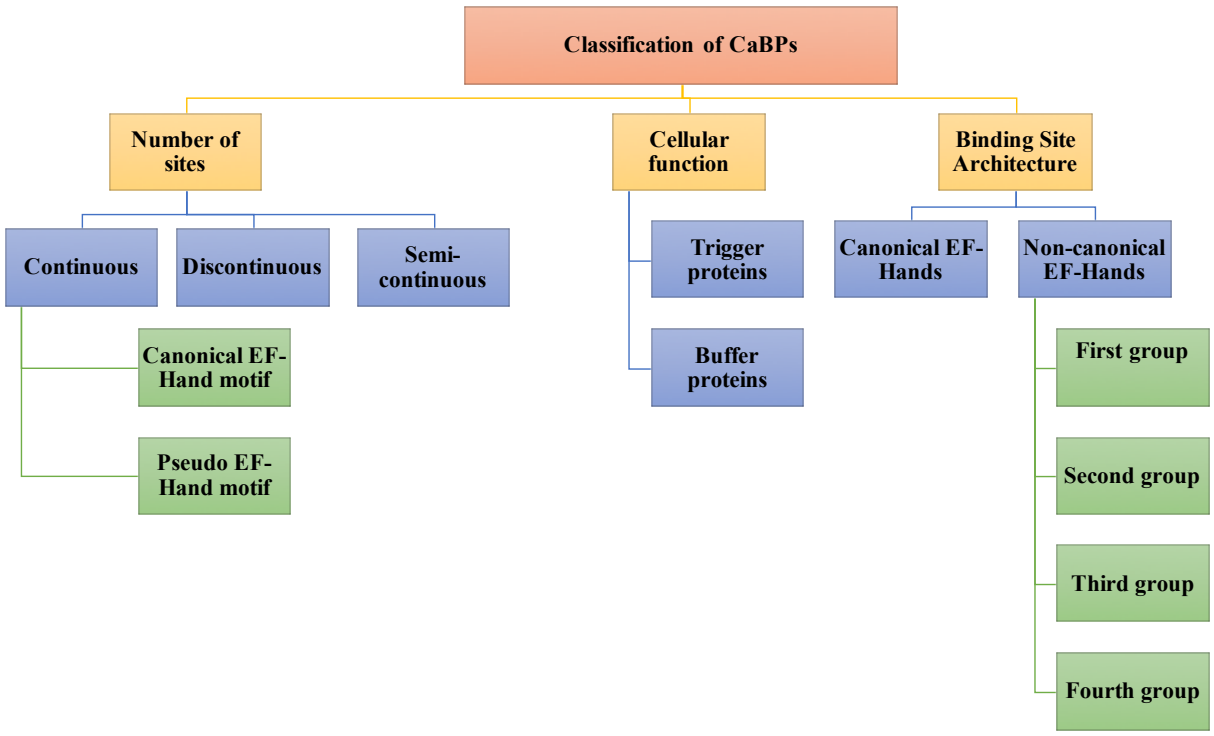


Figure 1.2 The schematic representation of the classification of calcium binding proteins on the basis of type of calcium binding site, function and the architecture of the binding site.

1.5 Classification of CaBPs on the basis of their "function"

In accordance with the role of the divalent ion in terms of functionality, we can categorize the calcium binding proteins into two groups: 1) trigger or sensor proteins e.g. calmodulin [5, 11, 12] and 2) buffer proteins such as parvalbumin [2, 13, 14] (Figure 1.2).

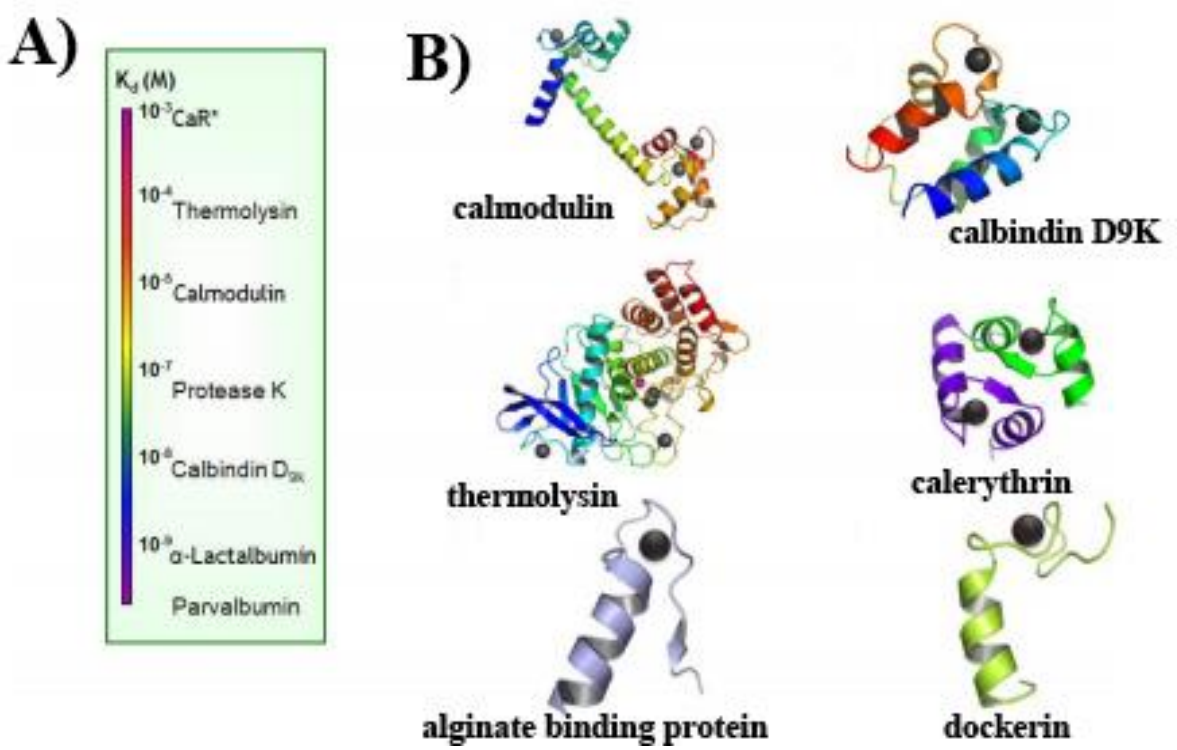


Figure 1.3 Classification of calcium-binding proteins on the basis of their calcium binding ability. The figure shows the different range Calcium binding affinities seen in different classes of *EF* hand binding domains. (Adapted from Jones *et. al* 2006 [7])

1.5.1 Calcium Sensor proteins

One of the most essential protein amongst the sensor proteins is calmodulin (CaM)[11, 15]. CaM is a relatively small (Molecular Weight: 16.7 KDa) intra-cellular protein. CaM has two domains

that show cooperative binding properties. The calcium binding to various CaM/ CaM-like proteins has been found to be associated with signal transduction via interacting with numerous proteins and regulating various important cellular processes. This signal transduction in sensor protein is positively correlated with calcium induced conformational change in these proteins which brings them to interact with various target proteins [16].

Although calcium binding does not always induce conformational changes (Judit *et al.*, 2005) as seen in many buffer proteins; but it is a prerequisite for many sensor proteins involved in various regulatory pathways. The sensor proteins as per their importance in most of the cellular processes have been widely studied. The conformational changes that Ca^{2+} brings in these structures are due to the chelation of Ca^{2+} with the interacting residues. The differences in the conformations were observed using X-ray crystallography and other biophysical techniques. The structural comparison of open and closed conformation suggests calcium-induced conformation changes were measured in the context of closed to open conformation of EF-hand motif after Ca^{2+} binding [15, 17, 18]. The differences in the structures are measured in the context of inter-helical angle. In the apo-protein these inter-helical angles are around ~130°-140°; on calcium binding to the motif the conformation of CaM changes to an open-state where the inter-helical angle is approximately ~ 90° [19] [18]. On conformational change or change in the inter-helical angle the hydrophobic residues in CaMs get exposed on the surface that enables it to interact with various target protein [17].

1.5.2 Calcium buffer proteins

Ca^{2+}-buffer proteins are a minor subset of the EF-hand protein family [20]. It has been seen that many of the EF-hand calcium binding sites in these proteins binds with very high affinity to Ca^{2+}. These proteins bind the free Ca^{2+}-ions to transmit the signal throughout the cell through different signalling pathways proteins such as Calbindin D_{9k} plays important role in possible removing harmful ions from the cytoplasm [21]. Many experimental structures of the proteins such as parvalumin and calbindin has been elucidated and deposited in both calcium-free and calcium-bound form [22]. The sensor proteins undergo large conformational changes upon Ca^{2+}-

binding compared to the buffer proteins that shows slight changes in the overall conformation of the molecule [20].

CaBPs such as Calretinin has six EF-hands containing Ca^{2+}-binding domains. It is predominantly expressed in certain neurons of the central and peripheral nervous system. However, these buffer proteins earlier thought to have simple roles are also involved in embryonic development and in mesothelioma cells. Interestingly, of the six EF-hand motifs, only 5 are functional; the 5th EF-hand shows low affinity and interacts with other binding partners of Calretinin and the top four domains show high cooperative binding which helps in the modulation of Ca^{2+}-signal[23].

1.6 Classification on the basis of the number of sites

In terms of the number of sites; the calcium binding proteins can be categorized into continuous, semi-continuous and discontinuous [24].

1.6.1 Continuous calcium binding sites

The continuous calcium binding site is formed by residues present in a protein sequence in a stretch, the EF-hand helix-loop-helix is one of the examples of continuous calcium binding site. The canonical EF-hand loop which binds to Ca^{2+} is formed by 12 residues stretch present in EF-hand motif [25]. In Class II sites or semi-continuous sites; one ligand coordinates from an amino acid sequence far from the rest of the binding sequence. Semi-continuous sites have been identified in the galactose binding protein, site I of subtillisin [26, 27] and site I of thermitase [28].

Interestingly, amongst all three classes of sites, the pentagonal bipyramidal geometry is most commonly utilized for Ca^{2+} binding. The divalent ion can bind with four to ten ligands in its primary coordination sphere, however it has been seen so far that it mostly binds with coordination sphere of six to seven ligands. Canonical EF-hand-motifs have most commonly a coordination sphere comprising of seven ligands. In this coordination, the Ca^{2+} ion interacts with seven oxygen ligands forming a pentagonal bipyramidal geometry [29].

EF-Hand proteins are extensively studied with more than 6000 EF-hand related entries in the NCBI Reference Sequences Data Bank. The continuous increase in the PDB entries of the CaBPs does reflect the biological importance of these proteins. Since the coining of the calcium binding EF-hand motif in 1973 by R. H. Kretsinger, the family of *EF*-hand proteins has extended to sixty-six subfamilies and more [30]. The continuous sites have two different patterns of calcium binding motifs canonical and non-canonical EF hand motifs.

1.6.2 Discontinuous calcium binding sites

The discontinuous binding sites are formed by residues present all over the protein (not in a stretch of consecutive residues as seen in most of binding sites) but in the three-dimensional arrangement they are in close proximity forming the calcium binding site [31]. The calcium binding site and its coordination in calcium binding varies from protein to protein implying that there are many ways that Ca^{2+} could bind to different proteins that could result in different binding affinities. The three Ca^{2+} ions share some of the same ligands but the binding sites are discontinuous [32].

1.7 Classification on the basis of the binding site architecture

The EF-hand motifs are branched into two major groups: Canonical EF-hands and Non-canonical EF-hands. The major difference in the two is the number of amino acids required to form the binding site.

1.7.1 *Canonical EF-hand Motif*

In the case of 12 residues canonical EF-hands such as calmodulin (CaM) which binds calcium primarily via side chain carboxylates or carbonyls (loop sequence positions 1, 3, 5, 12). One of the first examples for continuous binding site CaBP is the EF-hand motif. It is composed of a highly conserved loop and is flanked by two alpha helices (helix-loop-helix). It can be further divided into canonical or classic EF-hand motifs and pseudo-EF-hand motifs. The canonical EF-hand made up of a thirty-residue contiguous polypeptide containing two helixes, helix I (helix E), helix II (helix F) and a loop between these two helices where Ca^{2+}-ion bind[1, 25, 33] .

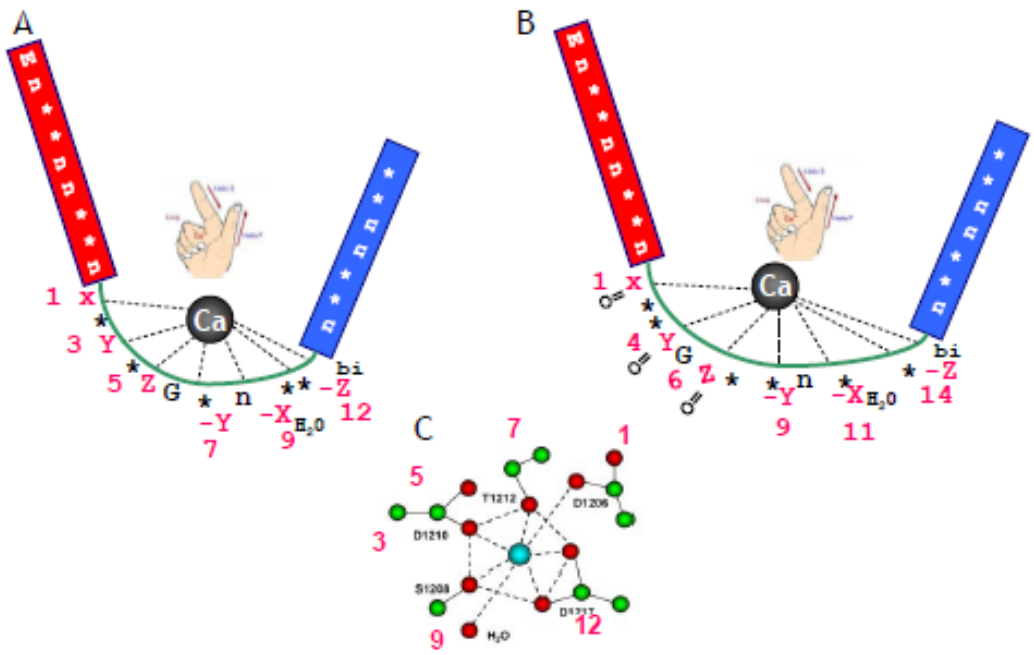


Figure 1.4 Schematics of the calcium binding EF-hand motif and the coordination of the calcium with the EF-hand motif (Adapted from EF-hand Wikipedia page).

The residue positioned at –X axis coordinates with the Ca^{2+} ion through a bridging water molecule. The EF-hand loop has a bidentate ligand (Glutamate or Aspartate) at –Z axis. On the other hand, in pseudo EF-hands, 14-residue EF-hand loop chelates calcium primarily via backbone carbonyls (positions 1, 4, 6, 9). In most cases the residue at the -X position coordinates with the Ca^{2+} ion via a bridged water molecule at the 9th position.

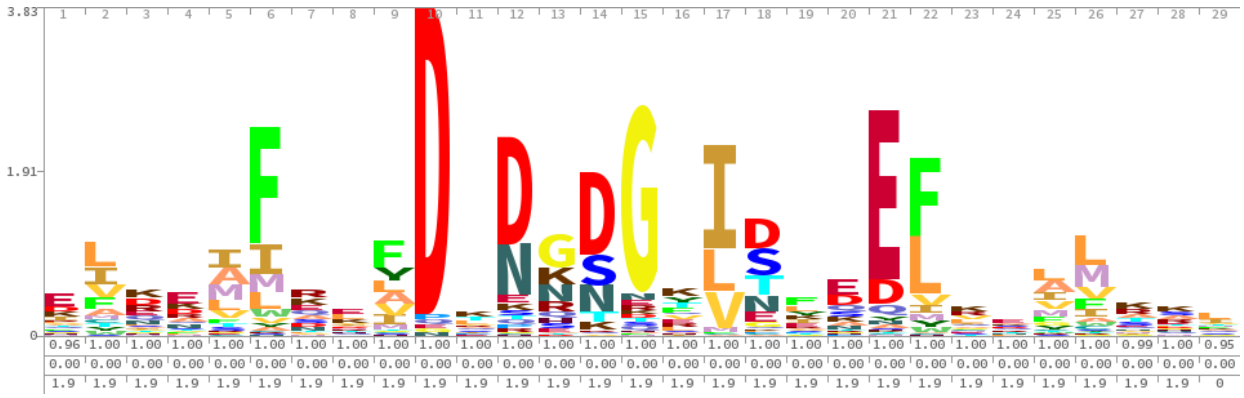


Figure 1.5 The Hidden Markov Model logo (HMM) of the calcium binding EF-hand motif showing the preference of amino acids at all the positions. The graphical representation shows the importance of specific residues at specific positions (Adapted from Wikipedia EF-hand page).

Out of the twenty-nine residues, the twelve residues which form loop are- 1, 3, 5, 7, 9 and 12 of the canonical binding loop that coordinate to the Ca^{2+}- ion and form a pentagonal bipyramidal array of six/seven oxygen ligands. Residues 1, 3 and 5 are contributors of the monodentate oxygen ligands via side chain oxygen atoms, usually carboxylate group of aspartate. Residue 12 is a bidentate oxygen ligand, a glutamate residue (92%) in most cases, which coordinates calcium via both side chain carboxylate oxygens. Residue 7 directly coordinates Ca^{2+} ion via its main chain oxygen whereas, residue 9 forms H-bonds to a water molecule that supplies the remaining Ca^{2+} ligand. This canonical motif is present in most of the EF-hand proteins.

The sequence analysis of the Ca^{2+}-ligand population revealed that the active-site has predominant acidic residues.

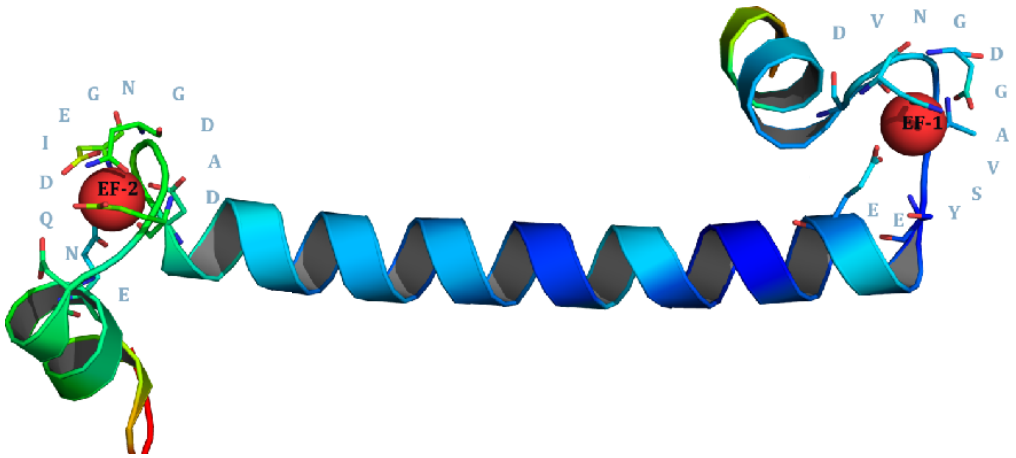


Figure 1.6 The EF-hand motif along with the loop residues are shown from *Eh*caBP1. The complete model is shown in cartoon representation and the residues in the EF-loop are shown in stick representation.

In the conventional canonical EF-hand loop the positions 1, 3 and 5 is highly conserved with aspartate present in 60% of the time at 1st position, 76 % at third position and 52% at the fifth position. One of the conventional EF-hand site is present in *Entamoeba Histolytica* calcium binding protein is shown in figure5. Besides having highly conserved positions the 2nd and 7th positions are the most variable [17].

1.7.2 Non-canonical EF-hand motif

The reasons to consider the EF-hand loop Ca^{2+} -binding as conventional binding is because of its wide presence in most of the organisms and crucial pathways. The structures of other helix-loop-helix CaBPs are also reported in which Ca^{2+} coordination varies from EF-hand called non-canonical or unconventional Ca^{2+} are binding motif. The non-canonical EF-loops are further classified into four groups.

1.7.2.11 First Group

The first type of non-canonical EF-hand motif does not require canonical ligands to bind Ca^{2+}-ion. This group contains two types of deviation. The second kind in this group are *EF*-hand loops that instead of binding via the side chains as seen in canonical EF-loops binds Ca^{2+} through an

increased use of main-chain carbonyl groups. The example of such mechanism is seen in EF4 of *Arabidopsis thaliana* Calcineurin B-like protein [34] (AtCBL2).

1.7.2.2 Second group

The second group of non-canonical EF-loops has two residues insertions in the Ca^{2+} binding loop also known as pseudo EF-loop (Ψ-hand) [17, 29]. This entire pattern is present in S100s and related proteins (Calbindin family). The Ψ-hand motif has a two residues insertion hence a modified coordination scheme is adapted. In the loop, residues positions 1, 4, 6 and 9 chelate the Ca^{2+}-ion via their main chain oxygen and residue position 11 chelates indirectly via bridging with a water molecule. The side chain carboxylate of residue position 14 provides the bidentate coordination. Similar to the *EF*-hand canonical loop, this terminal loop residue is almost always a glutamate (E).

1.7.2.3 Third group

The third types of non-canonical EF-hands are very rare. There has been only one report of a Ca^{2+} binding loop, shorter than the canonical Ca^{2+} binding EF-loop [35]. In this type of motif instead of 12 stretch of residues contributing to binding, the loop contains only 11 residues.

1.7.2.4 Fourth Group

Fourth type of Ca^{2+}-coordination was observed in the acellular true slime mold *Physarum polycephalum*. Mostly in the EF-hand loops the predominant type of molecular coordination is pentagonal bipyramidal geometry. In *Physarum polycephalum* the EF-hand motif is octahedral coordination [35]. In this coordination, the typical 12th position glutamate does not participate in calcium coordination and instead of one water molecule is involved in coordination of $Ca^{2+.}$

1.8. Other Calcium Binding Motifs

Since the identification of first ever calcium binding EF-hand motif in 1973, many other calcium binding motifs have been identified and reported in the literature. Some of the recently known calcium binding motifs are C2 domain motif, ANP-like domains and a new class of calcium binding motif known as calcium blades [36, 37]. There are four main conserved domains (C1-

C4) in the α, β and γ isoforms of mammalian calcium-dependent protein kinase C (PKC). Of this C2 domain is a Ca^{2+} -ion binding motif [38-40].

In contrast to the helix-loop-helix EF-hand structure, a new structural calcium binding motif has been discovered [41]. This motif shows deviation from typical EF-hands, and it includes helix loop-strand, helix-loop-turn, strand-loop helix, strand-loop-strand and many structural regions lacking a regular secondary structure element before or after the two kDxDxDG containing loop. These calcium binding loops effectively bind to Ca^{2+}ions and the calcium-binding ligands structurally align extremely well. These calcium binding motifs play regulatory role in various cellular processes and also communicate the simple regulatory signal into various functional responses. The exceptional versatility of the different calcium binding motifs is clearly reflected on the growing database of experimental structures and literature on these proteins that reveals a great diversity of conformations, domain organization and structural responses to calcium.

1.9. Calcium binding affinity and metal selectivity of proteins

CaBPs have varying 3D structures, Ca^{2+} binding affinities and various roles in the biological system. Many of the proteins families that binds Ca^{2+} have little in common besides their ability to bind Ca^{2+}-ions. The natural environment does puts certain limitations to the range of Ca^{2+} affinities which are compatible with the particular biological role of each protein. The serine binding protein, thermitase has an extremely strong binding affinity ($K_d = 10^{-10}$ M () while the binding affinity of the buffering protein Calbindin D9k is four orders of magnitude weaker ($K_d = 10^{-6}$ M) and the binding affinity of Concanavalin A is 7 orders of magnitude weaker ($K_d = 10^{-3}$ M) [42-46]

One of the key factor that influences the proper functioning is the selectivity of Ca^{2+} binding proteins over other physiologically relevant metals. The high level of intracellular Mg^{2+}, for example, as compared to Ca^{2+}, imposes the necessity of discrimination against Mg^{2+} for CaBPs operating inside the cell. In fact, Mg^{2+} is very similar to Ca^{2+} in that it is a divalent alkaline earth metal that favours oxygen ligands. However, an important difference is that the rate of water loss from the hydration shell of Mg^{2+} has a much slower rate [47], and its complexes are generally coordinated with six ligands, with water molecules occupying at least two sites, which makes Mg^{2+} unfavourable for binding to the highly coordinated irregular binding sites of CaBPs. It is

these kind of sites that produce the greatest structural change upon metal binding, enabling Ca^{2+} binding to act as a conformational trigger [48].

The high net negative charge of the ligands at most Ca^{2+} binding sites favours divalent and trivalent cation binding over monovalent ions like Na^{+}. In addition, the cavity size of a Ca^{2+} binding site affects the size range of ions than can be accommodated within it. The radius of Ca^{2+} is 1.0 Å, whereas other divalent ions Mn^{2+}, Fe^{2+}, Co^{2+}, Ni^{2+}, Cu^{2+}, Zn^{2+}, and Cd^{2+} have radii of 0.75, 0.70, 0.68, 0.65, 0.60, 0.65 and 0.90 Å respectively (Table 1.1). Cd^{2+} which has the closest radius to Ca^{2+} has a pr*EF*erential binding to strong sulphur containing proteins. Another factor in metal selectivity may be the free energy of metal-ion dehydration which varies with hydration number of the free ion and ionic radius [49]. The abundance of calcium binding residues is from mostly turn/loop like structure's. One of the most likely reason is due to the flexible nature of loop/turn and also the ability to supply a large number of bulky amino acids from a short stretch of protein sequence. Some of the motifs other than the *EF*-hand binds Ca^{2+} with different amino acid positioning and residues such as C2 domains [5, 9, 10, 31, 48].

1.10 Influence of the calcium binding affinity

The Ca^{2+} -ion concentration has a great influence over the its binding because of the different order of magnitude of calcium influx in the cell. The extracellular matrix has a higher level of Ca^{2+} -ion storage compared to the cytoplasm. The variation in the concentration is as high as 2mM in extracelluar matrix to 0.2µM in cytoplasm (~10000 fold). This calcium flux is governed during the essential physiological processes and has a critical role in tissue formations, differentiations in the cell structure and signal transduction. One of such example is Cadherin, a cell adhesion molecule [50].

Interestingly, following an alteration in the membrane potential or signal stimulation, the Ca^{2+} ion concentrations inside the cell can be elevated by more than a 100 fold. One of the driving forces for the calcium enabled signaling is due to the variation in Ca^{2+} -ion concentration which is also governed by different calcium binding affinities.

In response to the extracellular stimuli the Ca^{2+}-ion binds to the calcium binding signature sequences present in the trigger proteins such as Calmodulin (CaM)15]. The binding of calcium with its target proteins triggers numerous functions. CaM is a very important protein present

ubiquitously in almost all eukaryotes, bacteria as well as in viruses. The functional diversity of CaM also depends on the cellular environment that has different concentration of Ca^{2+}-ions different time points [49].

1.11 Concentration dependent action in calcium signalling

In typical the cellular ionization of the Ca^{2+} gradient follows in the following order:

> Extracellular space (~10^{-3}mM)

> Sarcoplasmic Reticulum (SR)/endoplasmic reticulum (ER) (~10^{-3} mM)

> cytosol (~10^{-7} M to ~10^{-5} µM)

> other internal calcium stores such as the mitochondrion (~10^{-7} M) and nucleus (~10^{-7} M).

1.12 Improper Calcium binding affinity causing diseases

Slightest variations in the sequences of calcium binding proteins can cause improper binding affinities in CaBPs can lead to disease such as osteoporosis, Alzheimer's and heart diseases. Some of the known disorders due to improper calcium binding affinities of the CaBPs are seen in Calmyrin, parathyroid calcium sensing receptor and Calreticulin which lead to the diseases mentioned b*EF*ore. Marfan syndrome is one of the rare diseases that is associated with mutation in Fibrillin-1 structure. It has a modular structure containing 47 epidermal growth factor-like (EGF-like) domains. This mutation in asparagine 2144 to serine (Asn2144Ser) in the calcium binding site causes decrease in calcium binding due to distortion in the alpha helix [51].

1.13 Summary of Calcium binding affinities of various EF containing proteins

Here we have summarized the calcium binding information from the published literature for the calcium binding affinities of *EF*-hand containing proteins along with their amino sequences that bind to the calcium.

Protein	Sequences/Canonical *EF* loops	Review of calcium binding affinity	Predicted Affinity	References
Bovine chains αα	I).DEDGDGEVDFQE	Contains low affinity calcium binding sites. The lower affinity calcium-binding sites titrated at a lower pH.	Low Affinity	[52]
Bovine chains αβ	I).DSDGDGECDFQE		Low Affinity	
Human chains β β	I).DNDGDGECDFQE	Six Ca^{2+}-binding sites which assumed to represent three for each β-monomer. Each β subunit was shown to bind one calcium ion with rather high affinity and two other calcium ions with lower affinity.	Low Affinity	[53]
Rat Chain β β	I).DEDGDGECDFQE	Rat brain S100b protein is characterized by two high-affinity Ca^{2+} binding sites with a KD of 2 X 10(-5) M and four lower affinity sites with KD about 10(-4) M.	Low Affinity	[54]
Frog pI 4-50 (FPV4- 50)	II).DQDKSGFIEEDE III).DSDGDGKIGVDE	Muscular parvalbumins from hake proteins have two high affinity sites	High Affinity High Affinity	[13]
Frog pl 4.88	I).DQDQSGFIEKEE II).DKDGDGKIGVDE	Parvalbumins exhibit two independent and equivalent high affinities Ca^{2+}-Mg^{2+} sites.	High Affinity High Affinity	[14]
Pike pl 5.00	I).DADASGFIEEEE	The intrinsic phenyl-alanine and tyrosine fluorescence of pike parvalbumins monitors the binding of Ca^{2+} ions to both their high affinity Ca^{2+} binding CD and *EF* sites.	High Affinity	[55]
Rabbit (RPV)	I).DKDKSGFIEEEE II).DKDGDGKIGADE I).DKDKSGFIEEDE II).DKDGDGKIGVEE	α-parvalbumins from rabbit exhibit two independent and equivalent high- affinity Ca^{2+}-Mg^{2+} sites.	High Affinity High Affinity High Affinity High Affinity	[14]

Rat (RTPV)		Parvalbumins: Each of their two functional sites binds Ca (II) with an affinity of about 10^8 M^{-1}.		[56]
Bovine cardics (BCTNC)	I)LGAEDGCISTKE II).DEDGSGTVDFDE III).DKNADGYIDLEE IV).DKNNDGRIDYDE	The C-terminal peptide contains two Ca2+-binding sites. The third and fourth sites in cardiac-muscle troponin C are represented by the so-called high-affinity Ca^{2+}/Mg^{2+}- binding sites.	Low Affinity Low Affinity High Affinity High Affinity	[57-59]
Amphioxus	I).DYNKDGSIQWED II).DINKDDVVSWEE III).DVSGDGIVDLEE	The two Amphioxus SCP's have three Ca-binding sites of high affinity: two calcium-specific ones and one Ca^{2+}-Mg^{2+} site.	Low Affinity Low Affinity Low Affinity	[60]
Nereis	I).DFDKDGAITRMD II).DTNEDNNISRDE III).DTNNDGLLSLEE	Ca2+ the three sites have the same intrinsic affinity (Ka = 1.7x108 M^{-1}) without co-operatively between the sites.	Low Affinity Low Affinity Low Affinity	[61]
Rabbit(RSLC2)	I).DQNRDGIIDKED II).DPEGKGTIKKQF	Myosin contains two DTNB light chains and binds 2 molecules of Ca (II) with high affinity.	Low Affinity Low Affinity	[62-65]
Scallop	I).DVDRDGFVSKDD	Concluded that both RLC-a and RLC-b bind only one Ca^{2+} with similar affinities to each other.	Low Affinity	[66]
Aequorin	I).DVNHNGKISLDE II).DKDQNGAITLDE III).DIDESGQLDVDE	The Kca for one of the two Ca^{2+} is approx. $7x10^6$ M^{-1}	Low Affinity High Affinity Low Affinity	[67, 68]

Calcineurin B	I).DLDNSGSLSVEE II).DTDGNGEVDFKE III).DMDKDGYISNGE IV).DKDGDGRISFEE	Demonstrate that Calcineurin is also a Ca^{2+}-binding protein with a high affinity for Ca^{2+} (10^{-6} M) in the presence of physio-logical concentrations of Mg^{2+}.	Low Affinity Low Affinity High Affinity High Affinity	[69]
Ca vector protein	I).DANGDGVIDFDE II).DEDGNGVIDIPE	CaVP binds 2 Ca^{2+} atoms in a non-cooperative way with intrinsic binding constant of $8.2x10^{6}$ M^{-1}forms a high affinity Ca^{2+}-dependent complex.	High Affinity High Affinity	[70]
F. Hepatica FH8	I).DRNGDGKVSAEE II).DKNKDGKLDLKE	FH8 displays low affinity for Ca^{2+}	Low Binder Low Binder	[71]
Human S100A	I).DANHDGRISFDE	Shows weak binding affinity for ca^{2+} One Ca^{2+}-binding site with micromolar affinity	Low Binder	[72]
Human Polycystin-2	I).DQDGDQELTEHE		Low Binder	[73]
Human Calnuc	I).DINSDGVLDEQE II).DTNQDRLVTLEE	Ca^{2+} binds with an affinity of 7μM and causes structural changes. They showed that Ca^{2+} binds to both sites with equal affinity.	Low Binder Low Binder	[74]
Human Centrin3	I).DTDKDEAIDYHE II).DDDDSGKISLRN III).DKDGDGEINQEE	Binds one Ca^{2+} with high and two Ca^{2+} with low affinity.	Low Binder Low Binder High Binder	[75]
Human Centrin2	I).DRDGDGEVSEQE	Binds only one Ca^{2+} per molecule with a significant affinity	High Binder	[76]

S. cerevisiae Centrin	I).DMNNDGFLDYHE II).DDDHTGKISIKN III).DLDGDDEINENE	Cdc31 has one high affinity Ca^{2+} - Mg^{2+} and two lower affinity Ca^{2+} sites.	Low Binder Low Binder High Binder	[77]
Human Calsenilin	I).DINKDGYITKEE II).DRNQDGVVTIEE	Affinities for Ca^{+2} binding at these two sites are greater than 1 µM.	High Binder High Binder	[78]_ENR*EF*_41

Protein	EF-Loop Prediction	K_a (M^{-1}) from the whole protein	Review of calcium binding affinity	Predicted Affinity	Ref.
Parvalbumin *Cyprinus carpio*	DQDKSGFIEEDE DSDGDGKIGVDE	K1= $2.7x10^9$ K2=$2.7x10^9$	The two metal sites of parvalbumin for Ca^{2+} with equilibrium constants of K_{Ca} = 2.7 X $10^{9}M^{-1}$	High Affinity High Affinity	*[79]*
Calmodulin *Bos taurus*	DKDGDGTITTKE DADGNGTIDFPE DKDGNGYISAAE DIDGDGQVNYEE	K1 = 1 x 10^7; K2 = 3.98 x 10^7; K3 = 3.16 x 10^6; K4 = 2.5 x 10^6	Calmodulin contains four relatively high affinity Ca^{2+}sites	High Affinity High Affinity High Affinity High Affinity	*[11]*
Caltractin *Chalmydomonas reinhardtii*	DTDGSGTIDAKE DKDGSGTIDFEE DDDNSGTITIKD DRNDDNEIDEDE	K1 = 8.30 x 10^5; K2 = 8.30 x 10^5, K3 = 6.25 x 10^3; K4 = 6.25 x 10^3	Ca^{2+} binding measurements demonstrated the binding of four Ca^{2+} ions to caltractin with two higher affinity and two lower affinity sites.	High Affinity High Affinity Low Affinity Low Affinity	*[80]* , [81]
Calmodulin-like protein *Homo sapiens*	DKDGDGCITTRE DRDGNGTVDFPE DKDGNGFVSAAE DTDGDGQVNYEE	K1=3.80 x 10^5, K2=1.90 x 10^5, K3=4.90 x 10^4, K4=1.20 x 10^4	Four Ca^{2+}-binding sites. Binding of the first two Ca^{2+}occurs with somewhat higher affinity than that of the last two Ca^{2+}.	High Affinity Low Affinity Low Affinity Low Affinity	*[12]*
Calbindin D9k *Bos taurus*	DKNGDGEVSFEE	K1 = 1.6 x 10^8, K2 = 4 x 10^8	Ca^{2+} ion binding to calbindin D9k wild type and with different set of mutants. (High Affinity)	Low Affinity	*[82]*
Calgranulin C *Sus scrofa*	DANQDEQVSFKE	K1=6.50 x 10^4	The protein binds one Ca^{2+}/monomer with a binding constant of about 2 x 10^4, a low affinity site	Low Affinity	*[83]*
GF14-loop1 *Arabidopsis*	ELDTLGEESYKD	K1=5.50 x 10^4	Low binding affinity exhibited by GF14 ω.	Low Affinity	*[84]*
Calhepatin *Lepidosiren paradoxa*	DKDKSGTLSVDE DTNKDGQVSWQE	K1=2.90 x 10^5 K2=6.00 x 10^3	The affinity constants determined agree with the fact that S100 protein affinity for Ca^{2+} is low, the affinity of the C-terminal *EF*-hand being greater than that of the N-terminal *EF*-hand.	Low Affinity Low Affinity	[85]

Table 1.1 & 1.2. The data is shown in a tabulated format where the first column represents the name of the protein and the organism followed by the amino acid sequence of the 12-mer that binds to the calcium. The third column represents the citations from the paper regarding the calcium binding abilities of these proteins. The fourth column is the classification on the basis of review of the binding affinities and the last column represents the author of the research article.

The review and the survey of so many CaBPs suggests that the calcium binding proteins with *EF* hand motifs plays very important roles in almost all the biological processes. Moreover, many CaBPs are extensive studies over the period of time; offering a rich database in the literature. This provides us with the opportunity to look for a a pattern to further build models for predictions.

1.14 References

1. Lewit-Bentley A, Réty S. EF-hand calcium-binding proteins. Current Opinion in Structural Biology. 2000;10(6):637-43. doi: http://dx.doi.org/10.1016/S0959-440X(00)00142-1.

2. Cates MS, Teodoro ML, Phillips GN. Molecular Mechanisms of Calcium and Magnesium Binding to Parvalbumin. Biophysical Journal. 2002;82(3):1133-46. doi: http://dx.doi.org/10.1016/S0006-3495(02)75472-6.

3. Ikura M. Calcium binding and conformational response in EF-hand proteins. Trends in Biochemical Sciences. 1996;21(1):14-7. doi: http://dx.doi.org/10.1016/S0968-0004(06)80021-6.

4. Palmer LC, Newcomb CJ, Kaltz SR, Spoerke ED, Stupp SI. Biomimetic Systems for Hydroxyapatite Mineralization Inspired By Bone and Enamel. Chemical reviews. 2008;108(11):4754-83. doi: 10.1021/cr8004422. PubMed PMID: PMC2593885.

5. Grabarek Z. Insights into modulation of calcium signaling by magnesium in calmodulin, troponin C and related EF-hand proteins. Biochimica et Biophysica Acta - Molecular Cell Research. 2011;1813(5):913-21. doi: 10.1016/j.bbamcr.2011.01.017.

6. Palta S, Saroa R, Palta A. Overview of the coagulation system. Indian Journal of Anaesthesia. 2014;58(5):515-23. doi: 10.4103/0019-5049.144643. PubMed PMID: PMC4260295.

7. Jones LM. Using Protein Design to Understand the Role of Electrostatic Interactions on Calcium Binding Affinity and

Molecular Recognition. Chemistry Dissertations. 2008;Paper 16.

8. Kim J, Kim HY. Functional analysis of a calcium-binding transcription factor involved in plant salt stress signaling. FEBS Letters. 2006;580(22):5251-6. doi: 10.1016/j.febslet.2006.08.050.

9. Berridge MJ, Lipp P, Bootman MD. The versatility and universality of calcium signalling. Nature Reviews Molecular Cell Biology. 2000;1(1):11-21.

10. Berridge MJ, Bootman MD, Roderick HL. Calcium signalling: Dynamics, homeostasis and remodelling. Nature Reviews Molecular Cell Biology. 2003;4(7):517-29. doi: 10.1038/nrm1155.

11. Linse S, Helmersson A, Forsen S. Calcium binding to calmodulin and its globular domains. J Biol Chem. 1991;266(13):8050-4. Epub 1991/05/05. PubMed PMID: 1902469.

12. Rhyner JA, Koller M, Durussel-Gerber I, Cox JA, Strehler EE. Characterization of the human calmodulin-like protein expressed in Escherichia coli. Biochemistry. 1992;31(51):12826-32. Epub 1992/12/29. PubMed PMID: 1334432.

13. Benzonana G, Capony JP, Pechere JF. The binding of calcium to muscular parvalbumins. Biochim Biophys Acta. 1972;278(1):110-6. Epub 1972/08/31. PubMed PMID: 4538395.

14. Haiech J, Derancourt J, Pechere JF, Demaille JG. Magnesium and calcium binding to parvalbumins: evidence for differences between parvalbumins and an explanation of their relaxing function. Biochemistry. 1979;18(13):2752-8. Epub 1979/06/26. PubMed PMID: 113029.

15. Babu YS, Bugg CE, Cook WJ. Structure of calmodulin refined at 2.2 A resolution. J Mol Biol. 1988;204(1):191-204. Epub 1988/11/05. PubMed PMID: 3145979.

16. Bhattacharya S, Bunick CG, Chazin WJ. Target selectivity in EF-hand calcium binding proteins. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research. 2004;1742(1):69-79. doi: http://dx.doi.org/10.1016/j.bbamcr.2004.09.002.

17. Gifford JL, Walsh MP, Vogel HJ. Structures and metal-ion-binding properties of the Ca2+-binding helix-loop-helix EF-hand motifs. Biochemical Journal. 2007;405(2):199-221. doi: 10.1042/BJ20070255.

18. Hoeflich KP, Ikura M. Calmodulin in action: diversity in target recognition and activation mechanisms. Cell. 2002;108(6):739-42. Epub 2002/04/17. PubMed PMID: 11955428.

19. Nelson DE, Rammesmayer G, Bohnert HJ. Regulation of cell-specific inositol metabolism and transport in plant salinity tolerance. The Plant cell. 1998;10(5):753-64. Epub 1998/06/03. PubMed PMID: 9596634; PubMed Central PMCID: PMCPMC144026.

20. Cates MS, Berry MB, Ho EL, Li Q, Potter JD, Phillips GN. Metal-ion affinity and specificity in EF-hand proteins: coordination geometry and domain plasticity in parvalbumin. Structure. 1999;7(10):1269-78. doi: http://dx.doi.org/10.1016/S0969-2126(00)80060-X.

21. Kiyota Y, Takeda-Shitaka M. Molecular Recognition Study on the Binding of Calcium to Calbindin D9k Based on 3D Reference Interaction Site Model Theory. The Journal of Physical Chemistry B. 2014;118(39):11496-503. doi: 10.1021/jp504822r.

22. Lee YH, Tanner JJ, Larson JD, Henzl MT. Crystal structure of a high-affinity variant of rat α-parvalbumin. Biochemistry. 2004;43(31):10008-17. doi: 10.1021/bi0492915.

23. Schwaller B. Calretinin: from a “simple” Ca2+ buffer to a multifunctional protein implicated in many biological processes. Frontiers in Neuroanatomy. 2014;8(3). doi: 10.3389/fnana.2014.00003.

24. Bindreither D, Lackner P. Structural diversity of calcium binding sites. General physiology and biophysics. 2009;28 Spec No Focus:F82-8. Epub 2010/01/23. PubMed PMID: 20093731.

25. Nakayama S, Moncrief ND, Kretsinger RH. Evolution of EF-hand calcium-modulated proteins. II. Domains of several subfamilies have diverse evolutionary histories. Journal of molecular evolution. 1992;34(5):416-48. Epub 1992/05/01. PubMed PMID: 1602495.

26. Wells JA, Ferrari E, Henner DJ, Estell DA, Chen EY. Cloning, sequencing, and secretion of Bacillus amyloliquefaciens subtillisin in Bacillus subtilis. Nucleic Acids Research. 1983;11(22):7911-25. doi: 10.1093/nar/11.22.7911.

27. Gilliland GL, Teplyakov A. Structural Calcium (Trypsin, Subtilisin). Encyclopedia of Inorganic and Bioinorganic Chemistry: John Wiley & Sons, Ltd; 2011.

28. Zeng J, Gao X, Dai Z, Tang B, Tang XF. Effects of metal ions on stability and activity of hyperthermophilic pyrolysin and further stabilization of this enzyme by modification of a Ca2+-binding site. Applied and environmental microbiology. 2014;80(9):2763-72. Epub 2014/02/25. doi: 10.1128/aem.00006-14. PubMed PMID: 24561589; PubMed Central PMCID: PMCPMC3993279.

29. Kirberger M, Wang X, Deng H, Yang W, Chen G, Yang JJ. Statistical analysis of structural characteristics of protein Ca2+-binding sites. JBIC Journal of Biological Inorganic Chemistry. 2008;13(7):1169-81. doi: 10.1007/s00775-008-0402-7.

30. Mazumder M, Padhan N, Bhattacharya A, Gourinath S. Prediction and Analysis of Canonical EF Hand Loop and Qualitative Estimation of Ca²⁺ Binding Affinity. PLoS ONE. 2014;9(4):e96202. doi: 10.1371/journal.pone.0096202.

31. Strynadka NCJ, James MNG. Crystal Structures of the Helix-Loop-Helix Calcium-Binding Proteins. Annual Review of Biochemistry. 1989;58(1):951-99. doi: 10.1146/annurev.bi.58.070189.004511.

32. Pedigo S, Shea MA. Discontinuous Equilibrium Titrations of Cooperative Calcium Binding to Calmodulin Monitored by 1-D 1H-Nuclear Magnetic Resonance Spectroscopy. Biochemistry. 1995;34(33):10676-89. doi: 10.1021/bi00033a044.

33. Strynadka NCJ, Cherney M, Sielecki AR, Li MX, Smillie LB, James MNG. Structural details of a calcium-induced molecular switch: X-ray crystallographic analysis of the calcium-saturated N-terminal domain of troponin C at 1.75 Å resolution11Edited by D. Rees. Journal of Molecular Biology. 1997;273(1):238-55. doi: http://dx.doi.org/10.1006/jmbi.1997.1257.

34. Luan S, Kudla J, Rodriguez-Concepcion M, Yalovsky S, Gruissem W. Calmodulins and Calcineurin B–like Proteins: Calcium Sensors for Specific Signal Response Coupling in Plants. The Plant cell. 2002;14(Suppl):s389-s400. doi: 10.1105/tpc.001115. PubMed PMID: PMC151268.

35. Debreczeni JE, Farkas L, Harmat V, Hetenyi C, Hajdu I, Zavodszky P, et al. Structural evidence for non-canonical binding of Ca2+ to a canonical EF-hand of a conventional myosin. J Biol Chem. 2005;280(50):41458-64. Epub 2005/10/18. doi: 10.1074/jbc.M506315200. PubMed PMID: 16227209.

36. Sutton RB, Davletov BA, Berghuis AM, Sudhof TC, Sprang SR. Structure of the first C2 domain of synaptotagmin I: A novel Ca2+/phospholipid-binding fold. Cell. 1995;80(6):929-38. doi: http://dx.doi.org/10.1016/0092-8674(95)90296-1.

37. Nalefski EA, Falke JJ. The C2 domain calcium-binding motif: Structural and functional diversity. Protein Science. 1996;5(12):2375-90. doi: 10.1002/pro.5560051201.

38. Kohout SC, Corbalán-García S, Torrecillas A, Goméz-Fernandéz JC, Falke JJ. C2 Domains of Protein Kinase C Isoforms α, β, and γ: Activation Parameters and Calcium Stoichiometries of the Membrane-Bound State. Biochemistry. 2002;41(38):11411-24. PubMed PMID: PMC3640336.

39. Gaertner TR, Kolodziej SJ, Wang D, Kobayashi R, Koomen JM, Stoops JK, et al. Comparative analyses of the three-dimensional structures and enzymatic properties of alpha, beta, gamma and delta isoforms of Ca2+-calmodulin-dependent protein kinase II. J Biol Chem. 2004;279(13):12484-94. Epub 2004/01/15. doi: 10.1074/jbc.M313597200. PubMed PMID: 14722083.

40. Cosentino-Gomes D, Rocco-Machado N, Meyer-Fernandes JR. Cell Signaling through Protein Kinase C Oxidation and Activation. International Journal of Molecular Sciences. 2012;13(9):10697-721. doi: 10.3390/ijms130910697. PubMed PMID: PMC3472709.

41. Rigden DJ, Galperin MY. The DxDxDG Motif for Calcium Binding: Multiple Structural Contexts and Implications for Evolution. Journal of Molecular Biology. 2004;343(4):971-84. doi: http://dx.doi.org/10.1016/j.jmb.2004.08.077.

42. Perona JJ, Craik CS. Structural basis of substrate specificity in the serine proteases. Protein Sci. 1995;4(3):337-60. Epub 1995/03/01. doi: 10.1002/pro.5560040301. PubMed PMID: 7795518; PubMed Central PMCID: PMCPMC2143081.

43. Linse S, Forsen S. Determinants that govern high-affinity calcium binding. Adv Second Messenger Phosphoprotein Res. 1995;30:89-151. Epub 1995/01/01. PubMed PMID: 7695999.

44. Dell'Orco D, Xue W-F, Thulin E, Linse S. Electrostatic Contributions to the Kinetics and Thermodynamics of Protein Assembly. Biophysical Journal. 2005;88(3):1991-2002. doi: 10.1529/biophysj.104.049189. PubMed PMID: PMC1305251.

45. el Aoumari A, Dupont E, Fromaget C, Jarry T, Briand JP, Kreitman B, et al. Immunolocalization of an extracellular domain of connexin43 in rat heart gap junctions. European journal of cell biology. 1991;56(2):391-400. Epub 1991/12/01. PubMed PMID: 1724962.

46. Bouckaert J, Loris R, Wyns L. Zinc/calcium- and cadmium/cadmium-substituted concanavalin A: interplay of metal binding, pH and molecular packing. Acta crystallographica Section D, Biological crystallography. 2000;56(Pt 12):1569-76. Epub 2000/11/28. PubMed PMID: 11092923.

47. Waluyo I, Huang C, Nordlund D, Bergmann U, Weiss TM, Pettersson LGM, et al. The structure of water in the hydration shell of cations from x-ray Raman and small angle x-ray scattering measurements. The Journal of Chemical Physics. 2011;134(6):064513. doi: 10.1063/1.3533958. PubMed PMID: PMC3188634.

48. McPhalen CA, Strynadka NC, James MN. Calcium-binding sites in proteins: a structural perspective. Advances in protein chemistry. 1991;42:77-144. Epub 1991/01/01. PubMed PMID: 1793008.

49. Williams RJP. The evolution of calcium biochemistry. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research. 2006;1763(11):1139-46. doi: http://dx.doi.org/10.1016/j.bbamcr.2006.08.042.

50. Gifford JL, Walsh MP, Vogel HJ. Structures and metal-ion-binding properties of the Ca2+-binding helix-loop-helix EF-hand motifs. Biochem J. 2007;405(2):199-221. Epub 2007/06/26. doi: BJ20070255 [pii]

10.1042/BJ20070255. PubMed PMID: 17590154.

51. Liu B, Lee RS, Biesiadecki BJ, Tikunova SB, Davis JP. Engineered troponin C constructs correct disease-related cardiac myofilament calcium sensitivity. J Biol Chem. 2012;287(24):20027-36. Epub 2012/04/19. doi: 10.1074/jbc.M111.334953. PubMed PMID: 22511780; PubMed Central PMCID: PMCPMC3370186.

52. Baudier J, Glasser N, Gerard D. Ions binding to S100 proteins. I. Calcium- and zinc-binding properties of bovine brain S100 alpha alpha, S100a (alpha beta), and S100b (beta beta) protein: Zn2+ regulates Ca2+ binding on S100b protein. J Biol Chem. 1986;261(18):8192-203. Epub 1986/06/25. PubMed PMID: 3722149.

53. Baudier J, Glasser N, Haglid K, Gerard D. Purification, characterization and ion binding properties of human brain S100b protein. Biochim Biophys Acta. 1984;790(2):164-73. Epub 1984/10/23. PubMed PMID: 6487634.

54. Baudier J, Labourdette G, Gerard D. Rat brain S100b protein: purification, characterization, and ion binding properties. A comparison with bovine S100b protein. Journal of neurochemistry. 1985;44(1):76-84. Epub 1985/01/01. PubMed PMID: 3964836.

55. Permyakov EA, Medvedkin VN, Kalinichenko LP, Burstein EA. Comparative study of physiochemical properties of two pike parvalbumins by means of their intrinsic tyrosyl and phenylalanyl fluorescence. Archives of biochemistry and biophysics. 1983;227(1):9-20. Epub 1983/11/01. PubMed PMID: 6639084.

56. Williams TC, Corson DC, Oikawa K, McCubbin WD, Kay CM, Sykes BD. 1H NMR spectroscopic studies of calcium-binding proteins. 3. Solution conformations of rat apo-alpha-parvalbumin and metal-bound rat alpha-parvalbumin. Biochemistry. 1986;25(7):1835-46. Epub 1986/04/08. PubMed PMID: 3707914.

57. Burtnick LD, Kay CM. The calcium-binding properties of bovine cardiac troponin C. FEBS Lett. 1977;75(1):105-10. Epub 1977/03/15. PubMed PMID: 852570.

58. Leavis PC, Kraft EL. Calcium binding to cardiac troponin C. Archives of biochemistry and biophysics. 1978;186(2):411-5. Epub 1978/03/01. PubMed PMID: 637570.

59. Barskaya NV, Gusev NB. Biological activities of bovine cardiac-muscle troponin C C-terminal peptide (residues 84-161). Biochem J. 1982;207(2):185-92. Epub 1982/11/01. PubMed PMID: 7159379; PubMed Central PMCID: PMC1153847.

60. Takagi T, Konishi K, Cox JA. Amino acid sequence of two sarcoplasmic calcium-binding proteins from the protochordate Amphioxus. Biochemistry. 1986;25(12):3585-92. doi: 10.1021/bi00360a017.

61. Cox JA, Stein EA. Characterization of a new sarcoplasmic calcium-binding protein with magnesium-induced cooperativity in the binding of calcium. Biochemistry. 1981;20(19):5430-6. Epub 1981/09/15. PubMed PMID: 7295685.

62. Bagshaw CR. On the location of the divalent metal binding sites and the light chain subunits of vertebrate myosin. Biochemistry. 1977;16(1):59-67. Epub 1977/01/11. PubMed PMID: 188447.

63. Sugden EA, Nihei T. The effects of calcium and magnesium ions on the adenosine triphosphatase and inosine triphosphatase activities of myosin A. Biochem J. 1969;113(5):821-7. Epub 1969/08/01. PubMed PMID: 4309596; PubMed Central PMCID: PMC1184772.

64. Okamoto Y, Yagi K. Inhibition by Mg2+ of the interaction of Ca2+ with spin-labeled g2 bound to myosin. Journal of biochemistry. 1977;82(3):835-7. Epub 1977/09/01. PubMed PMID: 199584.

65. Alexis MN, Gratzer WB. Interaction of skeletal myosin light chains with calcium ions. Biochemistry. 1978;17(12):2319-25. Epub 1978/06/13. PubMed PMID: 678511.

66. Morita F, Kondo S, Tomari K, Minowa O, Ikura M, Hikichi K. Calcium binding and conformation of regulatory light chains of smooth muscle myosin of scallop. Journal of biochemistry. 1985;97(2):553-61. Epub 1985/02/01. PubMed PMID: 4008468.

67. Shimomura O, Johnson FH. Calcium binding, quantum yield, and emitting molecule in aequorin bioluminescence. Nature. 1970;227(5265):1356-7. Epub 1970/09/26. PubMed PMID: 4393938.

68. Allen DG, Blinks JR, Prendergast FG. Aequorin luminescence: relation of light emission to calcium concentration--a calcium-independent component. Science. 1977;195(4282):996-8. Epub 1977/03/11. PubMed PMID: 841325.

69. Klee CB, Crouch TH, Krinks MH. Calcineurin: a calcium- and calmodulin-binding protein of the nervous system. Proc Natl Acad Sci U S A. 1979;76(12):6270-3. Epub 1979/12/01. PubMed PMID: 293720; PubMed Central PMCID: PMC411845.

70. Cox JA. Isolation and characterization of a new Mr 18,000 protein with calcium vector properties in amphioxus muscle and identification of its endogenous target protein. J Biol Chem. 1986;261(28):13173-8. Epub 1986/10/05. PubMed PMID: 3759955.

71. Fraga H, Faria TQ, Pinto F, Almeida A, Brito RM, Damas AM. FH8--a small EF-hand protein from Fasciola hepatica. FEBS J. 2010;277(24):5072-85. Epub 2010/11/17. doi: 10.1111/j.1742-4658.2010.07912.x. PubMed PMID: 21078120.

72. Babini E, Bertini I, Borsi V, Calderone V, Hu X, Luchinat C, et al. Structural characterization of human S100A16, a low-affinity calcium binder. J Biol Inorg Chem. 2011;16(2):243-56. Epub 2010/11/04. doi: 10.1007/s00775-010-0721-3. PubMed PMID: 21046186.

73. Celic A, Petri ET, Demeler B, Ehrlich BE, Boggon TJ. Domain mapping of the polycystin-2 C-terminal tail using de novo molecular modeling and biophysical analysis. J Biol Chem. 2008;283(42):28305-12. Epub 2008/08/13. doi: M802743200 [pii]

10.1074/jbc.M802743200. PubMed PMID: 18694932; PubMed Central PMCID: PMC2568934.

74. Kanuru M, Samuel JJ, Balivada LM, Aradhyam GK. Ion-binding properties of Calnuc, Ca2+ versus Mg2+--Calnuc adopts additional and unusual Ca2+-binding sites upon interaction with G-protein. FEBS J. 2009;276(9):2529-46. Epub 2009/03/24. doi: EJB6977 [pii]

10.1111/j.1742-4658.2009.06977.x. PubMed PMID: 19302560.

75. Cox JA, Tirone F, Durussel I, Firanescu C, Blouquit Y, Duchambon P, et al. Calcium and magnesium binding to human centrin 3 and interaction with target peptides. Biochemistry. 2005;44(3):840-50. Epub 2005/01/19. doi: 10.1021/bi048294e. PubMed PMID: 15654740.

76. Durussel I, Blouquit Y, Middendorp S, Craescu CT, Cox JA. Cation- and peptide-binding properties of human centrin 2. FEBS Lett. 2000;472(2-3):208-12. Epub 2000/05/02. doi: S0014-5793(00)01452-6 [pii]. PubMed PMID: 10788612.

77. Miron S, Durand D, Chilom C, Perez J, Craescu CT. Binding of calcium, magnesium, and target peptides to Cdc31, the centrin of yeast Saccharomyces cerevisiae. Biochemistry. 2011;50(29):6409-22. Epub 2011/07/01. doi: 10.1021/bi200518d. PubMed PMID: 21714500.

78. Yu L, Sun C, Mendoza R, Wang J, Matayoshi ED, Hebert E, et al. Solution structure and calcium-binding properties of EF-hands 3 and 4 of calsenilin. Protein Sci. 2007;16(11):2502-9. Epub 2007/10/27. doi: 16/11/2502 [pii]

10.1110/ps.072928007. PubMed PMID: 17962406; PubMed Central PMCID: PMC2211699.

79. Moeschler HJ, Schaer JJ, Cox JA. A thermodynamic analysis of the binding of calcium and magnesium ions to parvalbumin. Eur J Biochem. 1980;111(1):73-8. Epub 1980/10/01. PubMed PMID: 6777163.

80. Weber C, Lee VD, Chazin WJ, Huang B. High level expression in Escherichia coli and characterization of the EF-hand calcium-binding protein caltractin. J Biol Chem. 1994;269(22):15795-802. Epub 1994/06/03. PubMed PMID: 8195234.

81. Veeraraghavan S, Fagan PA, Hu H, Lee V, Harper JF, Huang B, et al. Structural independence of the two EF-hand domains of caltractin. J Biol Chem. 2002;277(32):28564-71. Epub 2002/05/30. doi: 10.1074/jbc.M112232200

M112232200 [pii]. PubMed PMID: 12034713.

82. Linse S, Johansson C, Brodin P, Grundstrom T, Drakenberg T, Forsen S. Electrostatic contributions to the binding of Ca2+ in calbindin D9k. Biochemistry. 1991;30(1):154-62. Epub 1991/01/08. PubMed PMID: 1988017.

83. Dell'Angelica EC, Schleicher CH, Santome JA. Primary structure and binding properties of calgranulin C, a novel S100-like calcium-binding protein from pig granulocytes. J Biol Chem. 1994;269(46):28929-36. Epub 1994/11/18. PubMed PMID: 7961855.

84. Lu G, Sehnke PC, Ferl RJ. Phosphorylation and calcium binding properties of an Arabidopsis GF14 brain protein homolog. The Plant cell. 1994;6(4):501-10. Epub 1994/04/01. doi: 10.1105/tpc.6.4.501

6/4/501 [pii]. PubMed PMID: 8205002; PubMed Central PMCID: PMC160453.

85. Di Pietro SM, Santome JA. Structural and biochemical characterization of calhepatin, an S100-like calcium-binding protein from the liver of lungfish (Lepidosiren paradoxa). Eur J Biochem. 2002;269(14):3433-41. Epub 2002/07/24. doi: 3023 [pii]. PubMed PMID: 12135482.

Chapter 2.

Prediction and Analysis of Canonical EF-hand Loop and Qualitative Estimation of Ca^{2+} Binding Affinity

2. 1 Abstract

Calcium signalling plays a very important role in almost all of the biological systems. Different Ca^{2+} binding proteins display different levels of binding affinities for Ca^{2+} ion. There are methods available to experimentally identify the binding affinity of small ions. Since it's not always possible to experimentally determine Ca^{2+} binding properties of EF-hand containing Calcium binding proteins (CaBPs), it is necessary to be able to predict this property from primary sequence using computational approach. The focus of this study was to annotate correctly canonical EF-hand motif and further classify these on the basis of their Ca^{2+} binding affinities using Support Vector Machine kernel classifiers. The canonical EF-hand loop sequences were taken from PDB to develop a precise and accurate classifier to classify Ca^{2+} binding loops with non-Ca^{2+} binding regions of EF-hand proteins. Using binary and amino acid composition features we achieved 100% accuracy through 5-fold cross validation. Next we proposed a novel *ab initio* method to predict the calcium binding affinity, where training datasets were generated on the basis of evolutionary information (PSSM scores). The best performing classifier with concatenated features of accessibility and hydrophilicity showed an accuracy of 87% on experimental test data set. Furthermore, we achieved 100% accuracy on an independent dataset obtained from recently published affinity observations. To investigate further, we performed a proteome wide prediction for *E. histolytica* and classified known EF-hand proteins, and found many probable Ca^{2+} binding sites. We compared our results with published pattern search method on *E. histolytica* proteome and demonstrated our method to be more specific and accurate for predicting potential canonical Ca^{2+} binding loops.

A web server CAL-EF-AFi based on the above approach is freely available at http://202.41.10.46/calb/index.html , all the datasets used in the study & proteome scan results are freely available at http://202.41.10.46/calb/dataset.html.

2.2 Introduction

Calcium ion signaling plays a major role in controlling most biological systems and many cellular functions, such as fertilization, motility, cell differentiation, proliferation and apoptosis, which are directly or indirectly regulated by Ca^{2+} [1]–[3]. In eukaryotes, there are elaborate mechanisms that are involved in maintaining Ca^{2+} homeostasis [4]. A defect in any of the components of the Ca^{2+} homeostasis/signalling system may have disastrous consequences including cell death. Recently, many CaBPs have also been identified in bacteria and viruses, raising the possibility that the prokaryotes may also have a Ca^{2+} regulatory system, particularly in relation to host-pathogen interactions [5], [6].

Ca^{2+} -ion is bound by a variety of proteins that are capable of binding with different affinities [7]–[9]. Such CaBPs can be classified into two categories, Ca^{2+} sensors and buffers. The major function of the first category of CaBPs is to sense the level of free intracellular $Ca^{2+}$and then to activate a suitable signaling pathway [10].

In general, CaBPs contain two well-defined Ca^{2+}-binding motifs: the EF-hand and C2 domains [11]. The EF-hand motif is the most frequently occurring Ca^{2+}-binding motif in eukaryotic systems [12]. There are more than 66 subfamilies [13] of EF–hand proteins and 3000 *EF*-hand related entries in the NCBI Data Bank [14]. An EF-hand is composed of a typical helix-loop-helix structural unit. This group is the largest and includes well-known members, such as calmodulin, troponin C and S100B. These proteins typically undergo a calcium-dependent conformational change which opens a target binding site [13]. Proteins, such as calbindin D9k do not undergo calcium-dependent conformational changes [15]–[17].

EF-hand motifs are divided into two major structural groups namely, the canonical EF-hands as seen in calmodulin (CaM) and prokaryotic CaM-like protein calerythrin, and the pseudo EF-hands exclusively found in the N-termini of S100 and S100-like proteins [18]. In either structural group, a pair of EF-hand motifs or pseudo EF-hand motifs forms a structural domain and is the

minimum requirement for Ca^{2+}-dependent activation. In general, one of the EF-hand motifs has a higher Ca^{2+}-binding affinity than the other. The canonical Ca^{2+}-binding loop is characterized by a sequence of 12 amino acid residues. In an EF-hand loop the Ca^{2+}- ion is coordinated in a pentagonal bipyramidal coordination. The six residues involved in the binding are in positions 1, 3, 5, 7, 9 and 12; these residues are denoted by X, Y, Z, -Y, -X and -Z.

In general, affinity constants of EF-hand domains for Ca^{2+} vary from micromolar (µM) to millimolar (mM), reflecting the diversity of functions carried out by these proteins in a range of Ca^{2+}-ion concentrations. There is an increase in stability and change in conformation upon Ca^{2+} -ion binding. Several residues found in an EF-hand loop are highly conserved and contribute to the stabilization and proper folding of the binding site. Factors such as biological, environmental, as well as the binding sequence have been shown to contribute to the calcium-binding affinity of these proteins [18]–[21].

A number of algorithms have been developed to computationally identify EF-hand containing CaBPs (EhCABPs) and Ca^{2+}-binding regions, including statistical, machine learning and pattern search approaches [22]–[24]. Recently, Franke *et al.* (2010) [24] proposed a method to estimate Ca^{2+}-binding affinity based on free energy calculations using crystal structures of CaBPs. However, this method has limited use due to unavailability of crystal structures in complex with calcium for large number of CaBPs. Moreover, no suitable method is available for the prediction of Ca^{2+}-binding affinity from primary sequence information. There was an early attempt by Boguta *et al* (1988) [25] to estimate the binding affinity of calcium for troponin C (TnC) superfamily proteins based on the prediction of the secondary structures. The results were convincing for some proteins which follow a typical TnC pattern [25] but not for any other protein family. Since it is not always possible to experimentally determine Ca^{2+}-binding properties of EF-hand-containing CaBPs, it is necessary to be able to predict this property from the primary sequence. In this report, we describe a method for computational prediction of Ca^{2+}-binding loops and their affinities for Ca^{2+} -ion from the amino acid sequences. This chapter describes approaches to find a better correlation of sequence to binding affinities in order to predict the sequence to function (association constant (K_a)) relationship. The results show that the tool (CAL-EF-AFi) described here is accurate and provides useful information about Ca^{2+}-binding properties to experimental biologists for both characterized and uncharacterized proteins.

2.3 Material and Methods

2.3.1 Expression, Purification and Preparation of Metal-free Protein Solutions

The data used in this study was generated by performing experiments on the systems available in our lab. Five different *Eh*CaBPs (*Eh*CaBP1, 3, 5, 6, and 7) were overexpressed and purified as described earlier in published paper by Rout *et al* [36] [37]. In order to obtain accurate measurements of Ca^{2+}-binding energetics, it was essential to have the protein in its apo-form with no contamination of Ca^{2+} in the buffers. Hence, all of the buffers used for isothermal titration calorimetry (ITC) were decalcified using Chelex 100 resin (Bio-Rad). Decalcified ITC buffer (100mM NaCl and 50mM Tris-Cl, pH 7.0) was prepared by treatment with Chelex 100 resin (Bio-Rad). Each protein solution was treated with 5mM EGTA and 2mM EDTA to remove Ca^{2+} and Mg^{2+}. The EDTA/EGTA bound to metal ions were removed from protein solution using Amicon ultra centrifugal filter devices (Millipore), through extensive buffer exchange (decalcified). Before the ITC experiment, the sample cell and injection syringe of the ITC machine (Microcal Inc.) were extensively cleaned using the decalcified buffer.

2.3.2 Isothermal Titration Calorimetry (ITC)

ITC experiments were performed on a MicroCal VP-ITC microcalorimeter at 25˚C. Samples were decalcified, centrifuged and degassed prior to titration. A typical titration consisted of injecting 2µl aliquots of 10–20mM $CaCl_2$ solution (diluted from 1 M standard $CaCl_2$ solution supplied by Sigma-Aldrich Chemicals) into 100–200 µM protein solution after every 3 minutes to ensure that the titration peak returned to the baseline prior to the next injection. A total of 70 injections were carried out. Aliquots of concentrated ligand solution were injected into the buffer solution (without the protein) in a separate ITC run, to subtract the heat of dilution. Two sets of titrations were carried out for each protein: (i) apo-*Eh*CaBP in 50mM Tris-Cl, pH 7.0 and 100mM NaCl and (ii) holo-*Eh*CaBP in 50mM Tris-Cl, pH 7.0 and 100mM NaCl. The ITC data were analyzed using the software ORIGIN (supplied with Omega Microcalorimeter). The amount of heat released per addition of the titrant was fitted to the best least squares model as given by Wiseman *et al.* (1989). For each titration, the stoichiometry (n), association constant (K_a) and enthalpy change (ΔH) were obtained directly from the ITC data, and the changes in

Gibbs free energy (ΔG), and entropy (ΔS) as well as the overall binding affinity or dissociation constant (K_d) were calculated according to equations a, b, and c.

$\Delta G = RT \ln K_a$ …….. (a)

$\Delta G = \Delta H - T\Delta S$…… (b)

$K_d = 1/K_a$ *or* $K_d = 1/\sqrt{K1K2K3}$…….. (c)

2.3.3 Dataset for EF loop predictions

To predict the presence of EF-hand loops and estimate their affinities for Ca^{2+}, the calcium-binding amino acid sequence pattern at PROSITE [38](http://prosite.expasy.org/PDOC00018) was used to retrieve sequences of the EF-hand family. In total 1379 different sequences were obtained. To further validate the reviewed sequences we used structures of proteins co-crystallized with $Ca^{2+}$from the Protein Data Bank [39] (PDB, http://www.rcsb.org/pdb/). In total, 1261 chains with EF-hand motifs were found. Once these sequences were downloaded, CD-HIT [40] was used to remove redundant sequences having more than 60% similarity. The PDB IDs are included in the Appendix I along with the sequences retrieved. We chose a relatively high value because the aim of the study was to identify the binding loop, which is a highly conserved 12-residue sequence. With less than a 60% threshold, the numbers of sequences available for classification were not sufficient. The sequence classifications were also carried out using thresholds of 90%, 70%, 60%, 50% of CD-HIT data is also shown in appendix I. Finally, a dataset of 100, 12-mer calcium-binding loop sequences for the positive training dataset (D1) as was generated. Similarly, a negative training dataset was built with 141 (D2), 12-mer sequences extracted from non-binding regions of EF-hand proteins. The datasets discussed in this chapter were alphanumerically numbered sequentially. The sequences used to build the Datasets D1 and D2 are shown in appendix I.

2.3.4 Dataset for binding affinity predictions

To develop a good classifier, the utmost requirement is to have a good dataset. For the estimation of binding affinity, a novel method was developed on the basis of PSSM score pattern in which calcium-binding loops were classified into two groups. Based on the correlation obtained between the PSSM scores and experimental binding affinity (Figure 1) a positive dataset with high PSSM scores (D3) (>5) consisting of 144, 12-mer sequences and a negative dataset (D4) with low PSSM scores (<5) containing 124 sequences were generated using the sequences obtained from PROSITE [38]. The sequences used to build the Datasets D3 and D4 are listed in appendix I.

To test the proposed model based on PSSM scores we used 19 EF loop sequences for which binding affinities were known from the literature (appendix I) as Test dataset (D5). To evaluate the performance of this classifier on a dataset that has not been used for training and testing, an independent dataset (D6) of binding affinity observations was obtained from Boguta *et al* (1988) [25] and recently published literature. After removing redundant EF-loop sequences, 50 unique sequences were obtained from recently published data and the K_a values listed in Boguta *et al* (1988) [25]. Furthermore, to check the performance and reliability of the classifier, we chose to perform ITC experiments on available *Eh*CaBPs, to test our predictions on the datasets obtained from literature. We were able to obtain K_a values of EhCaBP1, 3, 5, 6, and 7; in total we listed affinities for 11 sites used here as a validation set (D7). The details of ITC experiments and results are also provided in datasets section on appendix I as D5, D6 and D7 with their experimental binding affinities classified on the basis of a thorough review of published papers that reported the binding constants. The classification details with supportive binding constants are listed under “Author’s Note” in Tables S2–S4 in appendix I.

2.3.5 Statistical Analysis

The expected (Exp) frequencies of amino acid residues were calculated from the average residue usage from the 1379 different sequences obtained from PROSITE [38]. The expected frequency for an amino acid residue of type A at position I will be Exp =(*NA/N) M*, where *NA*=total number of amino acid residues of type A in the analyzed set of sequences, excluding position *i*, *N*=total number of all amino acid residues in the analyzed set of sequences, excluding position *i*, and *M*=total number of sequences, i.e., the sum of *i*th positions in the analyzed set of sequences. The expected frequencies for residues were calculated similarly. For each amino acid residue at a given position, the deviation of the observed (Obs) values from the Exp values was estimated by the χ^2 criterion according to the formula (Obs – Exp) ^{2}/Exp. For each residue or codon, the χ^2 value was estimated separately with one degree of freedom. The sums of all 20 (61) χ^2 values for each residue (codon) at the given position gave the total deviation for the given position with 19 (60) degrees of freedom. To evaluate the range of differences between the C-terminal regions and the neighboring fragments, a pairwise comparison between them was performed. For this purpose, each position in the sequence was treated as a set containing 20 groups of data and the difference between them was calculated by the χ^2 criterion using the following formula:

$$\sum_{i=1}^{K}[(m_i / M - n_i / N)^2 MN / (m_i + n_i)]$$

where m_i and n_i are frequencies of amino acid residues in the two positions of the sequence under comparison, M and N are total numbers of amino acid residues in the compared positions, and K is equal to 20 because each position may be occupied by any of 20 different amino acids. At a significance level <0.001, Obs was considered to be different from Exp if the χ^2 exceeded 10.8, 43.8 and 99.6 for one, 19 and 60 degrees of freedom, respectively.

2.3.6 Generation of a position-specific scoring matrix

In this study, a simple position-specific scoring matrix (PSSM) was generated from the amino acid composition (AAC) of the calcium-binding loops in canonical EF-hands. The standard amino acid frequencies, which show how often each residue was found in each site in the binding loop, was taken from Marsden *et al.*, 1990 [41]. In this matrix, every column can be interpreted as a discrete probability distribution of the amino acid residues at that position and the values in the matrix can be inferred as probabilities of a given amino acid occurring at a given position. Therefore, for a sequence of length **m**, the product of the relative frequencies from the matrix corresponding to each amino acid in each position of the sequence is the probability of discovering such a sequence in the EF-hand loop. We generated two different scoring matrices, one with simple relative frequency of amino acids and the other with log likelihood frequency for the PSSM [42]–[44]. The log ratio matrix was generated using equations 1 and 2.

$$Sij = q + bPi / n + b \qquad (1)$$

$$Msij = \log(Sij / Pi) \qquad (2)$$

Where Sij is the probability of amino acid i at position j in matrix S, q is the observed counts of amino acid type i at position j, Pi is the probability of amino acid type i, b is the pseudo count which is considered here as square root of the total number of training sequences and n is the number of training sequences. In equation (2) Msij represents the foreground model (representing true homology) and Pi is the background model (chance that a match occurs at random). The background probability or the chance of amino acid match occurrence at random was calculated using the BLOSUM62 substitution matrix [45].

2.3.7 Support Vector Machine training for classification

SVM is a machine learning tool that is being extensively used for classification and optimization of complex problems. It is particularly attractive to biological sequence analysis due to its ability to handle noise, large datasets, large input spaces and high variability [46], [47]. In this study all of the SVM models have been developed using libSVM [48]. Parameter selection was carried out using grid search so that the classifier can accurately predict unknown test data from the model. In the radial basis function (RBF) kernel, there are two parameters, C and g, but it is not known *a priori* what values of these two parameters are best for a given problem [48]. To obtain the best parameters, a grid search was carried out using cross validation. A Perl check once spelling script was written in-house to check combinations of features in an iterative manner using CUDA based libSVM [49]. A descriptive flowchart of the feature selection algorithm is provided in Figure S4 in appendix I.

2.3.8 Five-fold cross-validation

A standard five-fold cross-validation technique was used to evaluate the performance of models, where the data set was randomly divided into five sets. The classifier was trained on four sets and the performance was assessed on the remaining fifth set. The process was repeated five times so that each set could be used once for testing. Finally, the average of the five sets was calculated as the measure of the performance of the classifier.

2.3.9 SVM model using binary and amino acid composition features

In this method, a Perl program was written to generate a window with 12 amino acids for negative and positive patterns. These sequence patterns were converted into binary patterns, where a pattern of length L was represented by a vector of dimension L×21 and each amino acid in that pattern was represented by a 21-feature vector (e.g. Asp by 1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,X) containing 20 amino acids and a dummy X. Each sequence of twelve amino acids was represented by 252 input vectors during model generation. The binary profile has been used in a number of existing methods [50], [51]. The second feature

used was AAC with an input vector of 20 X 12 dimensions. AAC is the fractional occurrence of each amino acid in the protein sequence.

$$Fi = Total\, number\, of\, Amino\, acid / Length\, of\, the\, protein$$

Where i can be any of the amino acids.

2.3.10 Feature extraction and model generation for binding affinity estimation

It has been observed in different studies [52], [53] that SVM performs well when combinations of two or more features are used as input vectors. Hence, hybrid models have been developed using one or more combinations of features. After testing combination of features using CUDA-based libSVM [49] the best performing features were used for developing various SVM models. Feature selection was carried out by scanning amino acid indices and by performing 5-fold cross validation using the in-house CUDA script. The four best performing amino acid properties used further for analysis were net charge [54](CC), hydrophobicity [55] (HYC), hydrophilicity [56] (HC) and accessibility [57] (AC) which were thus used for further analysis. Only the better performing models (AC&CC, AC&HC, AC&HYC, AC&HC&HYC, and AC&HYC&CC), which use combinations of the four best performing amino acid properties, are discussed in this study.

2.3.11 Classifier performance metrics

The performance of our method was computed and tested using the following figures of merit. As mentioned above, the performance has been evaluated by five-fold cross validation as follows:

1) **Sensitivity** (or recall) is the coverage of positives i.e. the percent of correctly predicted Ca^{2+}-binding 12-mers and correct estimation of their affinity.

$$Sensitivity = [TP/(TP + FN)] \times 100$$

2) **Specificity** is the coverage of negatives, that is, the percent of correctly predicted Ca^{2+} non-binding 12-mers and correct estimation of their affinity.

$$Specificity = [TN/(TN + FP)] \times 100$$

3) **Accuracy** is the percentage of correctly predicted positives and negatives.

$$Accuracy = [(TP + TN)/(TP + FP + TN + FN)] \times 100$$

4) **MCC** - Matthews's correlation coefficient is the statistical parameter to assess the quality of the prediction and account for unbalancing in data(Matthews 1975). An MCC equal to 1 is regarded as a perfect prediction, whereas that equal to 0 indicates a completely random prediction.

$$MCC = (TP)(TN) - (FP)(FN) / \sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}$$

[TP = true positive; FN = false negative; TN = true negative; FP = false positive]

5) **AUC** (Area under the ROC Curve) - Receiver Operating Curve (ROC) and AUC were computed using SPSS software. It generates ROC curves and calculates AUC by ranking the decision values.

[TP =true positive; FN =false negative; TN =true negative; FP =false positive]

2.4 Results

A few experimental methods based on biophysical techniques such as Isothermal titration calorimetry (ITC) surface plasmon resonance (SPR) & fluorescence [26] are available for determination of Ca^{2+}-binding parameters. However, these are expensive and time consuming. To the best of our knowledge, no prediction method has been developed so far that can be used to estimate Ca^{2+}-binding properties of a protein from primary sequence. Therefore, a comprehensive study was carried out first to identify Ca^{2+}-binding EF loops and then their Ca^{2+}-binding affinities. In this study, we have constructed two support vector machines (SVM), one for prediction of loop regions and the other for estimation of binding affinity.

2.4.1 Position-specific scoring matrix

After obtaining position-specific scoring matrix (PSSM) scores using equations (1) and (2) (described in Materials Methods) for all the sequences obtained from the literature, we calculated the correlation coefficient between the experimental affinity constants (K_a) and PSSM to be 0.61 (Figure 1). While this correlation is clearly positive, it was not possible to classify the affinity of all the sequences solely using PSSM scores. Therefore, a systematic attempt was made to first

predict the presence of canonical EF-hand loops from amino acid sequence and then estimate the binding affinities qualitatively based on evolutionary information using SVMs.

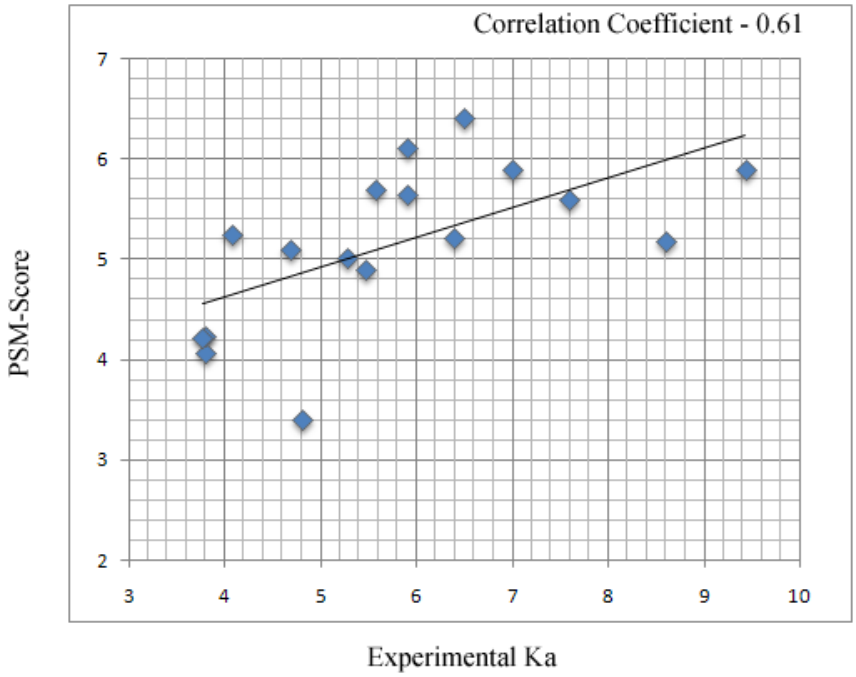


Figure 2.1 a) Plot of affinity vs. PSSM for the test data set (D5).The calculated correlation obtained was 0.61 using (Marsden, Shaw *et al*. 1990) amino acid frequencies.

2.4.2 Amino acid composition distinguishes Ca^{2+}-binding and non-binding regions

A statistical analysis was carried out to determine which amino acids are found unusually frequently in EF-hand motif sequences using the entire PFAM EF-hand database. Glycine, glutamic acid, asparagine and especially aspartate were determined to occur more frequently in Ca^{2+}-binding loop regions than in non-binding regions at a 99.9% confidence level. Alanine, phenylalanine, leucine and especially methionine were overrepresented in non-binding regions (Figure 2). The relative frequency of amino acids at each position is listed in Table S1 in

appendix I. The analysis suggested that EF-hand Ca^{2+}-binding loops have a specific amino acid composition, and that it is possible to identify these loops from the primary sequence.

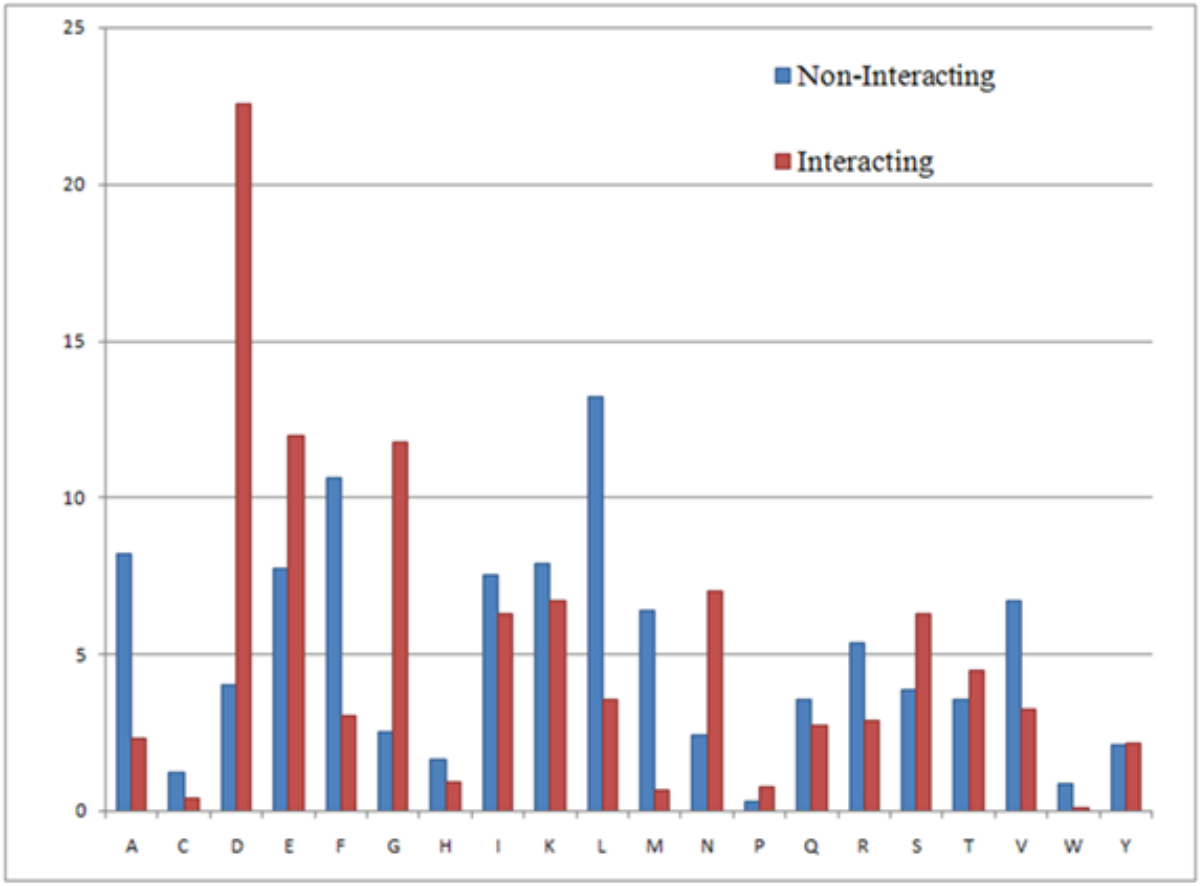


Figure 2.2. Amino acid composition of the 12-mer long Ca^{2+}-binding region ("Interacting") and the non-binding region ("Non-Interacting") of EF-hand proteins.

2.4.3 Experimental determination of Ca^{2+}-binding properties of EhCaBPs

In order to validate the theoretical predictions, experiments were carried out to determine qualitative and quantitative aspects of the affinity of some *Eh*CaBPs for Ca^{2+}. Ca^{2+}-binding properties of these proteins were tested by $^{45}Ca^{2+}$ overlay assay on western blotted pure recombinant *Eh*CaBP1, 3, 5, 6, and 7 proteins. All of these proteins were found to bind $^{45}Ca^{2+}$ as observed by autoradiography (data not shown). ITC was used to determine the molar

stoichiometry of the binding of the cations to these *Eh*CaBPs, as well as the binding constants and associated thermodynamic parameters (Table 2.1).

Table 2.1. Summary of macroscopic binding constants and thermodynamic parameters obtained from the ITC studies of Ca^{2+}-binding isotherm of *Eh*CaBPs at 25°C.

Ligand	Titrant	No. of experimental Ca^{2+}-binding sites (n)	K_a (M^{-1})	K_d	ΔH (cal/mol)	ΔS (cal/mol)	ΔG (kcal/mol)
Ca^{2+}	***Eh*CaBP1**	4	K1=5.25x10^3±4.0x10^2	130.72 μM	-1860±0	10.8	-4.84
			K2=1.41x10^4±9.5x10^2		2.3x10^5±0	790	-4.6x10^2
			K3=5.10x10^5±2.8x10^4		2.4x10^5±1.82x10^3	-780	-7.56
			K4=1.55x10^6±7.3x10^4		-7981±1.86x10^3	1.56	-8.44
	***Eh*CaBP3**	2	K1=4.00x10^6±5.3x10^5	1.85 μM	-1.605x10^4±86.6	-23.6	-9.0
			K2=7.28x10^4±5.3x10^3		-7573±10^4	-3.16	-6.63
	***Eh*CaBP5**	2	K=1.18x10^7±1.47x10^6	85 nM	-1.84x10^4±61.79	-29.4	-9.64
	***Eh*CaBP6**	2	K1=1.07x10^5±1.1x10^4	46 μM	702±17.6	25.4	-6.86
			K2=4.44x10^3±1.1x10^2		5244±45.9	34.3	-4.97
	***Eh*CaBP7**	2	K1=1.04x10^6±2.5x10^5	3.12 μM	-1807±96.5	21.5	-8.2
			K2=9.86x10^4±6.8x10^3		-5413±96.5	4.69	-6.81

The sequences and binding affinities of these proteins were used in the validation dataset (D7) for validation of the classifier's efficiency on experimental data. The data plotted after ITC experiments are shown in the Figure 2.3.

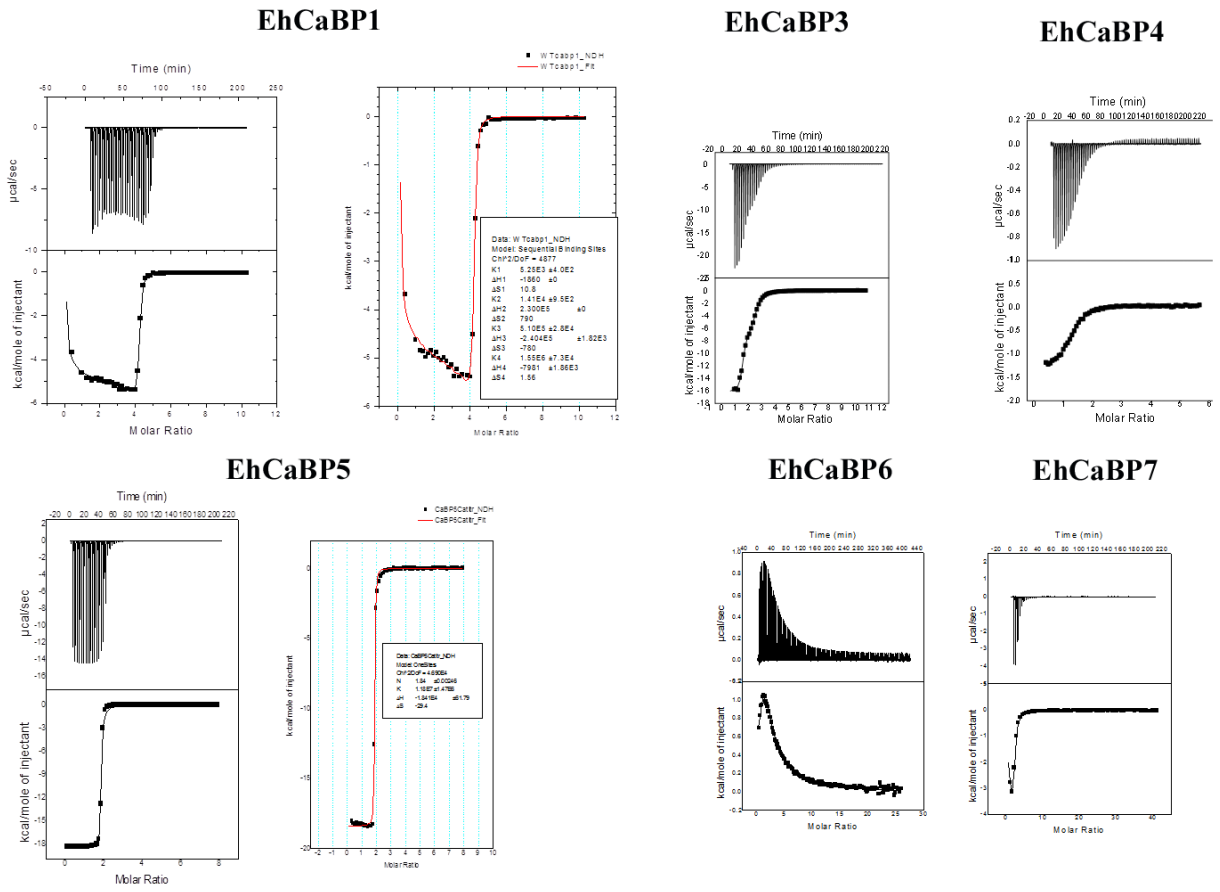


Figure 2.3 Isothermal titration calorimetric analysis of Ca^{2+}-binding to apo-*Eh*CaBPs**.** ITC experiments were carried out as described under "Materials and methods". Plot of kcal mol^{-1} of heat absorbed/released per injection of $CaCl_2$ as a function of molar ratio of Ca^{2+}: protein at 25°C is shown. For all titrations, the top panels represent the raw data (power: time) and the bottom panels represent integrated binding isotherms. The solid line represents the best nonlinear fit to the experimental data. Binding isotherm for A: *Eh*CaBP3; B: *Eh*CaBP4; C: *Eh*CaBP5; D: *Eh*CaBP6 and E: *Eh*CaBP7. Thermodynamic parameters obtained are summarized in the Table 1.

2.4.4 SVM models predict the presence of EF loop regions

Two different models were generated using both binary pattern and amino acid composition (AAC) for loop identification. Both AAC and binary pattern were calculated, and used as input for classification of Ca^{2+}-binding EF-hand loops and non-Ca^{2+}-binding 12-mers in EF-hand proteins using SVM. The models were generated by using different types of kernels, such as polynomial, radial basis function (RBF) and linear. The performance of each kernel function was evaluated by five-fold cross validation. During model generation, the RBF kernel showed the best results.

The RBF kernel function using binary and AAC standalone features most accurately predicted the presence of EF-loop regions. An accuracy of 100% was achieved with D1 and D2. The remarkable performance of binary and AAC is due to the high conservation of sequence and structure among EF-hand loops that have been used in this study. Normally, the default threshold value (0) was used for the SVM classifier to discriminate between Ca^{2+}-binding EF-hand loops and non-Ca^{2+}-binding 12-mers in EF-hand proteins. The sites with a prediction score close to 1 are most likely to be an EF-hand calcium-binding loop region. All performance measures and the learning parameters for the RBF kernel are listed in Table 2.2.

Table 2.2 **The Performance of SVM Models with different learning parameters on D1 and D2 dataset**. Using binary patterns and AA (amino acid) composition [γ **(g)** (in RBF kernel), c: parameter for trade-off between training error & margin] where SN–sensitivity, SP–specificity, ACC-accuracy, MCC–Matthews Correlation Coefficient.

Features	**C**	**G**	**SN**	**SP**	**ACC**	**MCC**
Binary	8	0.008	100	100	100	1
AA	0.125	0.008	100	100	100	1

2.4.5 Accessibility and hydrophilic (AC&HC)-based classifier provide the best estimation of binding affinity. Various SVM models using a combination of features were developed to

estimate the affinity of Ca^{2+} for the EF-hand loop. The predictions of binding constants were not as accurate as the predictions of EF-hand loops due to the limited availability of experimental data on binding constants and the high level of diversity in amino acid sequence with relation to binding affinity.

In this study, we have developed a position-specific scoring matrix for EF-hand loop regions and scored (equation [1] and [2]) the sequences from the annotated data set using Perl scripts developed in-house. Based on the PSSM scores, we classified high (D3) and low (D4) binding groups for the 12-mer region to train the classifier. The binding constants, obtained from the literature (Table S2 in appendix I) and data obtained from ITC studies of *Eh*CaBPs were used as the test dataset and validation dataset (Table S3 in appendix I) respectively. Since it is generally believed that different physico-chemical properties contribute to the structure and function of protein sequences, these properties should also contribute to Ca^{2+}-binding affinity. Therefore, we have developed several SVM models (data not shown) to achieve better accuracy using combinations of several amino acid features, and have obtained the different physio-chemical properties using the amino acid index database (http://www.genome.jp/aaindex/). Only the best performing models are discussed here.

For the 24-dimension input vectors consisting of accessibility (AC) and charge (CC), the values of sensitivity, specificity and accuracy were 90.97, 87.10, 90.30 and 90.91, 75.00, 84.21 for training and test datasets respectively. We were also able to achieve a Matthews's correlation coefficient (MCC) of 0.78 for the training datasets (D3 & D4) and 0.67 for the test (D5) dataset.

The classifier consisting of concatenated features of accessibility (AC) and hydrophilic (HC) scores showed the best performance when tested on the training and the test datasets, achieving an MCC of 0.87 and 0.81 and an accuracy of 94.78 and 89.47 for D3–D4 and D5 datasets, respectively. The superior performance of this classifier compared to other hybrid models is also indicated by its values for sensitivity and specificity of 95.83 and 91.00 respectively for the training dataset, and 81.82 and 100.0 respectively for the test dataset.

Several other hybrid models (AC&CC, AC&HC&HYC, AC&HYC&CC and AC&HYC) were also generated with amino acid features-based classifiers; however, their performances were not

better than the AC&HC-based classifier. The list of figures of merit of all the classifiers used can be found in Tables 2.3 and 2.4.

Table 2.3 **The Performance of SVM Models on PSSM based training dataset D3 & D4.**

The Performance of SVM Models on PSSM based training dataset D3 & D4 with different learning parameters on various hybrid models [γ (g) (in RBF kernel), c: parameter for trade-off between training error & margin] where SN–sensitivity, SP–specificity, ACC-accuracy, MCC–Matthews Correlation Coefficient, AUC/ROC-Area under curve/ Receiver Operating Curve.

Features	**C**	**g**	**SN**	**SP**	**ACC**	**MCC**	**AUC / ROC**
AC&CC	32768	0	90.97	87.1	90.30	0.78	0.94
AC&HC	***8***	***0.03***	***95.83***	***91.0***	***94.78***	***0.87***	**0.97**
AC&HC&HYC	2	0.13	94.44	91.0	94.78	0.86	0.97
AC&HYC&CC	2048	0	91.67	90.32	91.42	0.82	0.96
AC&HYC	2048	0	91.67	88.7	91.04	0.8	0.95

Table 2.4 **The Performance of SVM Models on test dataset D5**. The Performance of SVM Models on test dataset D5 (experimental binding affinities obtained from literature) with different learning parameters.

Features	**SN**	**SP**	**ACC**	**MCC**
AC&CC	90.91	75.00	84.21	0.67
AC&HC	**81.82**	**100**	**89.47**	**0.81**
AC&HC&HYC	72.73	87.50	78.95	0.6
AC&HYC&CC	90.91	75.00	84.21	0.67

AC&HYC	90.91	75.00	84.21	0.67

The quality of the performance of the AC & HC-based classifier is also indicated by receiver operating characteristic (ROC) plots, which we computed for all the models discussed in this study. ROC is commonly used to evaluate the discrimination ability of a classifier. If the area under the ROC curve is larger, it means the classifier has better discrimination ability.

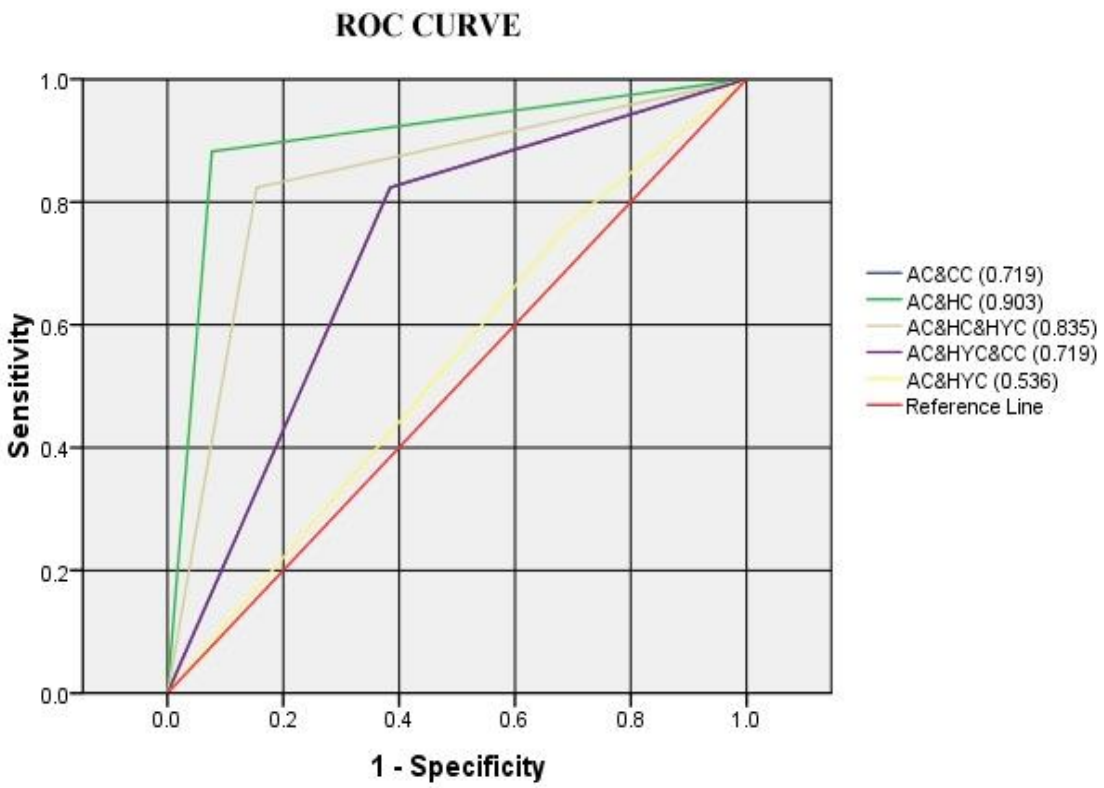


Figure 2.4 ROC plot of the best performing SVM classifiers. ROC plot of AC&CC, AC&HC, AC&HC&HYC, AC&HYC&CC and AC&HYC for the datasets D5-D7 set. Receiver operating characteristic (ROC) plot used for depicting relative trade-offs between true positive and false positives. The corresponding AUC value of each model is shown in brackets.

We were able to achieve an AUC of 0.97 with the training dataset and 0.903 with the experimental datasets (D5 & D7) using the AC&HC-based classifier (Figure S4 in Appendix I).

2.4.7 Prediction of Ca^{2+}binding of an independent dataset

After obtaining the best performing model, it was important to evaluate the performance of this classifier on a dataset that has not been used for training and testing. In order to check the unbiased prediction efficiency of the model, in addition to the test dataset, an independent dataset (D6) with 35 unique troponin C superfamily binding sites (Boguta *et al* 1988) and 15 unique sites (Table S4 in Appendix I) were tested using our classifier. The classifier predicted 21 high binders (true positives), 19 low binders (true negatives), and 10 high binders (false negatives) that were predicted as low binding sites. When using the diverse datasets and binding affinities obtained from different researchers working under different experimental conditions, the overall accuracy achieved was 80.0%.

2.4.7 The validation dataset

The performance of AC and HC-based classifier was even better when tested on the experimentally obtained binding affinities from EhCaBPs. We achieved an accuracy of 90.91 and MCC of 0.83. The performances of other classifiers for the validation dataset D7 are listed in table 2.5.

Table 2.5 **The Performance of SVM Models on validation dataset with experimentally derived binding affinity from EhCaBPs (D7)**.The Performance of SVM Models on validation dataset with experimentally derived binding affinity from EhCaBPs (D7)with different learning parameters on various hybrid models [γ (g) (in RBF kernel), c: parameter for trade-off between training error & margin] where SN–sensitivity, SP–specificity, ACC-accuracy, MCC–Matthews Correlation Co*EF*ficient, AUC/ROC-Area under curve/ Receiver Operating Curve.

Features	**SN**	**SP**	**ACC**	**MCC**
AC&CC	83.33	60	72.73	0.45
AC&HC	**100**	**80**	**90.91**	**0.83**
AC&HC&HYC	83.33	80	81.82	0.63

AC&HYC&CC	83.33	60	72.73	0.45
AC&HYC	66.67	60	63.64	0.27

2.4.8 E. histolytica proteome analysis: Computational prediction of Ca^{2+}-binding properties of EhCaBPs

In this section, we used 'CAL-*EF*-AFi' to scan the *E. histolytica* proteome in order to predict all Ca^{2+}-binding canonical EF-hand loops in this organism. A previous computational study [27] showed that there are 27 CaBPs containing EF-hand motifs present in *E. histolytica*. Our scanning results picked all the known canonical EF hands with more than one EF-hand loop region. Apart from the sequences used in the test dataset (Ehcabp1, 3, 5–7); we also predicted the relative affinities of other *Eh*CaBPs (8–27). In total, we predicted 36 Ca^{2+}-binding sites (Table S5 in Appendix I) out of which 24 were predicted to be low-affinity sequences and the remaining 12 sites were predicted to have high affinity for Ca^{2+}.

2.4.9 Comparison with existing methods

The performance of the classifier was compared with PFAM based HMM profile search and Calpred [28] on the *E. histolytica* proteome. In light of earlier bioinformatics studies by Bhattacharya *et al.* and availability of *E. histolytica* strain HM-1: IMSS for wet lab experiments, we chose the *E .histolytica* proteome for comparison. Although this is not a benchmark dataset, it was important to validate our classifier's accuracy to find EF-hand containing Ca^{2+}-binding sites in large databases and proteomes. A total of 41 EF-hand protein sequences were predicted using the pattern search method whereas CAL-EF-AFi found 58 probable sequences with 153 binding loops.

Based on the results obtained by PFAM pattern search, few of the predictions with high threshold values (Table S6 in appendix I) appear to be false positives. The tertiary structures of

all these proteins have not been determined yet, but they lacked the number of amino acids required to form a typical EF-hand structural motif. Similarly, we scanned *Eh*CaBPs with Calpred (using all the modules available), which identified EF-hand proteins but predicted false positives; all the residues in the full-length protein sequence were predicted as calcium binding (site). To investigate further we used sequences with known structures (D1 & D2) in Calpred and found similar false-positive predictions here as well. A thorough analysis (Table S6 in appendix I) of the results from different methods for the identification of EF-hand Ca^{2+}-binding sites suggests that the method proposed here to be most suitable for prediction of Ca^{2+}-binding sites and relative affinity constants and is also useful for whole proteome scans.

A schematic representation for the data input, algorithm implementation and experimental strategy overview is shown in Figure 2.5.

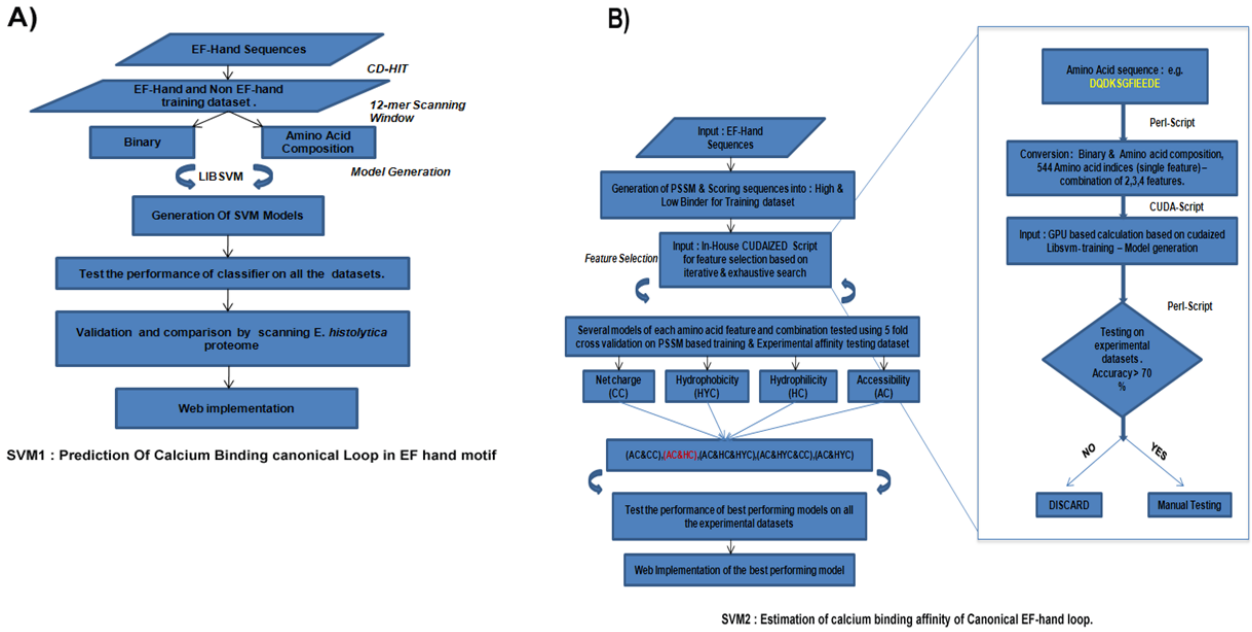


Figure 2.5 Schematic representation of the procedure for model development and feature selection for EF-hand loop region prediction and estimation of binding affinity and its web implementation. The procedure is explained in details in the "Methods" section. A). A group of

sequences with known EF-hand structural motif were downloaded and further classified into two groups after removing the redundancy using CD-HIT. The sequences were further converted into binary and amino acid composition (AAC) profile for SVM input. Models were generated using LIBSVM and were tested on all the datasets (D3-D6) and further validated by comparing and scanning E. *histolytica* proteome. B). Non-redundant sequences of EF-hand loops from known structures were classified into two groups on the basis of scores obtained from position specific scoring metrics. The sequences were then converted into binary, AAC and different amino acid indices patterns. We have generated both standalone and combination of features (2, 3, 4, 5) using in house Perl script. The input vectors were trained using LIBSVM and cudized LIBSVM and selected on the basis of their performance on experimental dataset using 5- fold cross validation accuracy threshold > 70 %. The best performing models selected from screening were further validated using three different experimentally derived datasets on *EF* hand motifs. The final step involved web implementation of the best (AC&HC) model.

2.4.10 Availability

CAL-*EF*-AFi is available at http://202.41.10.46/calb/index.html and all the datasets used in the study as well as the proteome scan results are available at http://202.41.10.46/calb/dataset.html

2.4.11 User Interface: Webserver

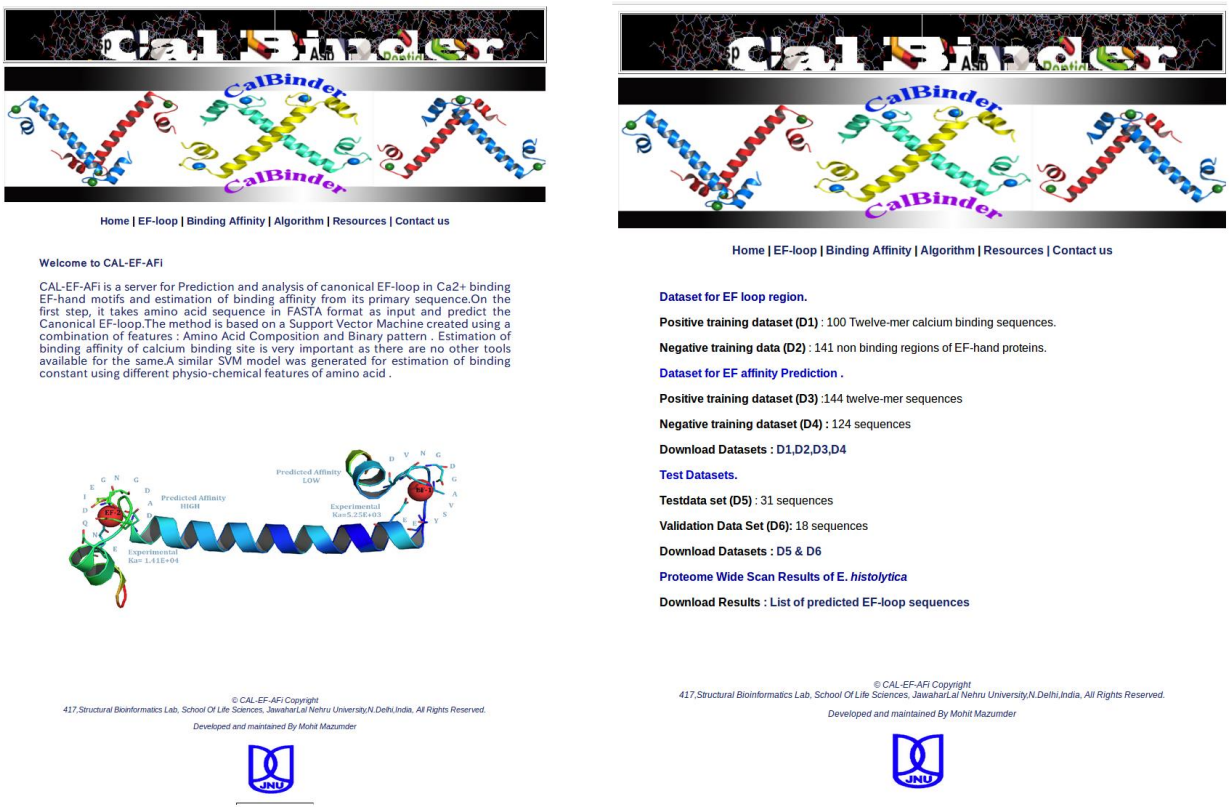


Figure 2.6. Screenshots of the home and resources page from Cal-*EF*-AFi webserver version 1.0.

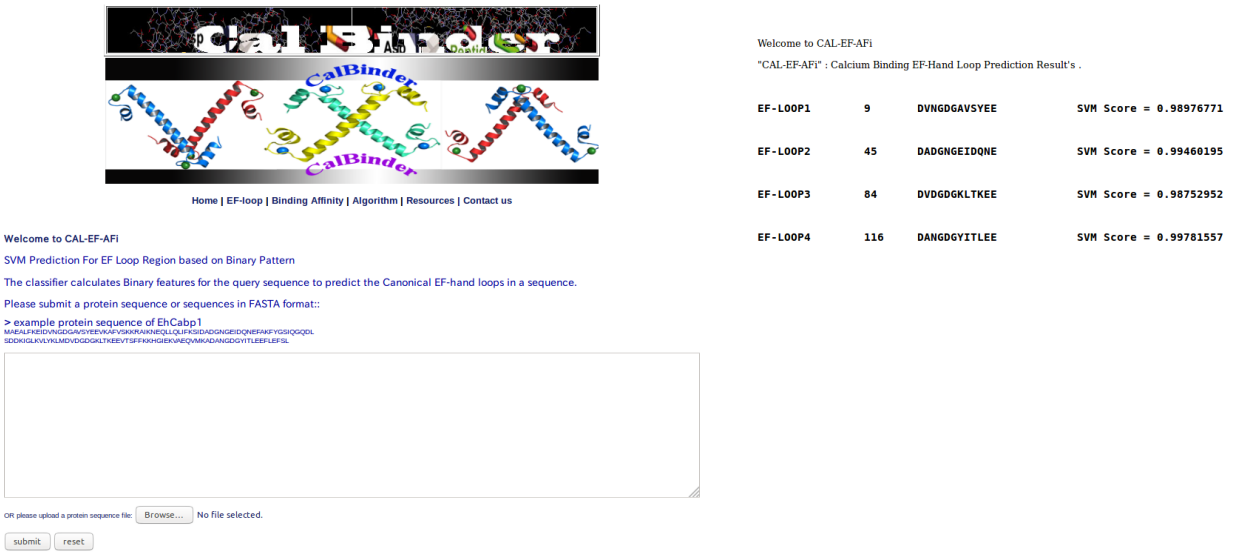


Figure 2.7. Screenshots of the binding loop prediction modules and result page from Cal-*EF*-AFi webserver version 1.0.

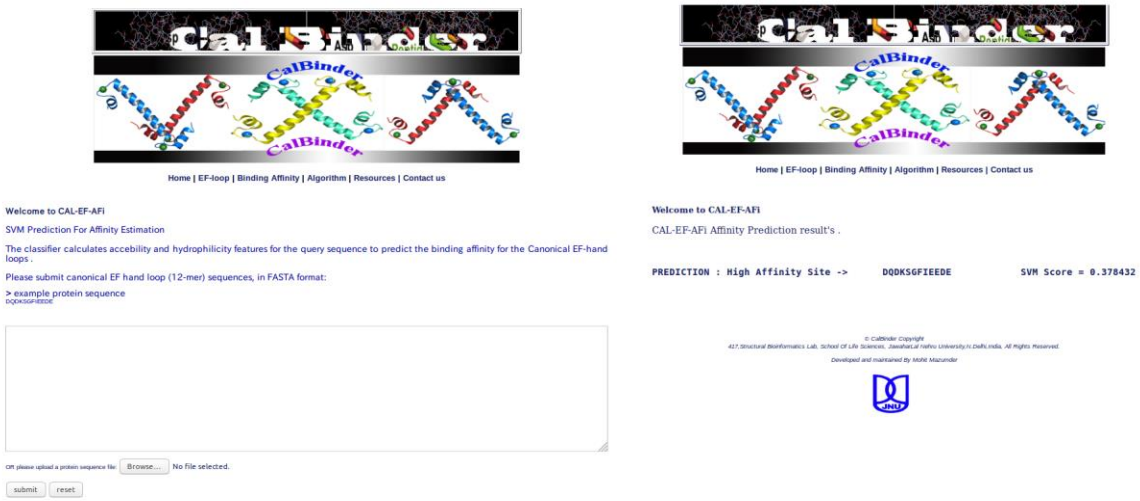


Figure 2.8. Screenshots of the binding affinity prediction modules and result page of affinity prediction from Cal-EF-AFi webserver version 1.0.

2.5 Discussion

In the current era of high-throughput next generation sequencing, where a large amount of genomic data is generated each day, prediction of gene functions and detailed annotation have become key aspects of computational genomics. The focus of this study is to annotate Ca^{2+}-binding EF-hand motif-containing proteins and further classify these on the basis of their Ca^{2+}-binding affinities.

Different Ca^{2+}-binding proteins display different levels of affinities for Ca^{2+}. The functions of these proteins in general depend on their affinity constants for Ca^{2+}. Ca^{2+}-sensor proteins such as calmodulin (CaM) display higher Ca^{2+} -binding affinities for their C-terminal domains than for their N-terminal domains [29]. Ca^{2+}-buffer proteins, such as parvalbumin have high binding affinity [30] and there is little or no change in their conformation upon binding Ca^{2+}. Hence, it is possible to predict the probable function of the proteins from Ca^{2+}-binding properties.

Many computational methods have been developed ever since identification of the first EF-hand domain as an approach for prediction of Ca^{2+}-binding sites. These methods were based on similarity search, energy based calculations, Bayesian statistical methods, machine learning approaches and graph theory [22], [31]–[33], where the input is either a primary amino acid sequence or a three-dimensional structure. A comparison of CAL-EF-AFi with the existing methods for identifying Ca^{2+}-binding sites is not suitable due to the dissimilarity in the prediction methods, input type and the datasets. One of the recently published machine learning approaches [28] to identify the calcium-binding region showed poor performance when compared with CAL-*EF*-AFi using a dataset of experimentally determined values. Some of the other methods, such as CaPS uses pattern search where EF-hand motif and Ca^{2+}-binding loops are predicted on the basis of patterns generated using a Hidden Markov Model (HMM) based on multiple sequence alignment of known EF-hand proteins. None of these methods, however, were able to predict the binding affinity of the identified Ca^{2+}-binding motifs. We trained the classifier using the sequences of EF hand motif binding and non-binding regions so that it could identify the Ca^{2+}-binding region in the EF-hand motif.

The performance of the classifier was also tested by analyzing the complete proteome of *E. histolytica*. Based on the scan results we found all of the reported Ca^{2+}-binding proteins, and also

identified new probable Ca^{2+}-binding sites. Our tool appeared to give better results in terms of identification of CaBPs as it identified more proteins including all known CaBPs. Other methods, such as PFAM-based HMM profile search and Calpred showed a significant number of false predictions. Our results, using all of the sequences in the test (D5) affinity estimation data set, suggest that the PSSM scores and experimental binding affinities are broadly correlated. In our study, we have classified proteins on the basis of relative binding affinity for Ca^{2+} in a semi-quantitative manner. There are a number of reasons that a precise quantitative analysis is still intractable. For one, a 12-mer motif alone does not determine the affinity since there may be contributions from other parts of the protein. Also, there was a cooperative involvement of more than one EF-hand loop in the binding of Ca^{2+}. This may be particularly important as a pair of EF-hands occur together [14]. Two EF-hand motifs in a pair (with very few exceptions) are related by an approximate two-fold rotational axis, forming a hydrophobic cavity opening which is likely to influence the binding affinity. Since these properties are difficult to factor in a model, our efforts are limited to classification of high and low binders rather than predicting precise binding affinities.

Our initial datasets contained 19 binding sites with experimental binding affinity data. In order to circumvent the problems associated with limited data, we generated training datasets based on the evolutionary information (PSSM) scores. A similar approach, where artificial datasets have been used in SVM, has been successful in greatly improving predictions [34], [35]. In these studies, researchers mainly generated negative datasets artificially for SVM classification. Our test data set with 19 sequences, independent dataset with 50 sequences and the validation data set with 11 sequences representing experimentally determined affinity data have shown extremely good results.

The results from the test and validation datasets, which includes relative affinities of several *EF*-hand proteins, suggest that our proposed model based on the PSSM method for estimation of binding affinity can help researchers to predict site-specific binding affinity. Experimental determination of such binding affinity is a limiting factor in Ca^{2+}-binding proteins because of the expense involved and time required carrying out the experiments. As mentioned above, the successful performance of the model with regards to prediction and estimation is attributed to the

accurate training of the classifier on a small number of training examples and the use of PSSM generated datasets.

2.6 Conclusion

CAL-EF-AFi can therefore be used to accurately and precisely scan proteomes of organisms for potential Ca^{2+}-binding sites of EF-hand proteins and estimate their probable relative binding affinities. Given the success of our classifier on the *E. histolytica* proteome scan, we expect its wider use in analyzing proteomes of other organisms.

In conclusion, we have developed a unique method, CAL-EF-AFi for identification and estimation of Ca^{2+}-binding sites and relative affinity. The program requires only the protein sequence for the prediction without prior knowledge of structural or biochemical information. The results predicted by the theoretical model were validated by experimental studies. Variation from the EF-hand consensus sequence can be used to predict qualitative Ca^{2+}-binding features. However, this may not be sufficient to understand the overall characteristics of CaBPs. The EF-hand motifs assemble to form a lobe (one partner affects the binding affinity of the other) and the Mg^{2+} affinities are not considered in this work due to limitation of experimental data available to date. Future plans include developing an even better algorithm with more information available from the literature. We hope that an increase in the availability of experimental data will help generate a more robust model.

2.7 References

1. Berridge MJ, Bootman MD, Lipp P (1998) Calcium--a life and death signal. Nature 395: 64648.
2. Ermak G, Davies KJ (2002) Calcium and oxidative stress: from cell signaling to cell death. Mol Immunol 38: 713-721.
3. Verkhratsky A (2007) Calcium and cell death. Subcell Biochem 45: 465-480.
4. Bencina M, Bagar T, Lah L, Krasevec N (2009) A comparative genomic analysis of calcium and proton signaling/homeostasis in Aspergillus species. Fungal Genet Biol 46 Suppl 1: S93-S104.
5. Gangola P, Rosen BP (1987) Maintenance of intracellular calcium in Escherichia coli. J Biol Chem 262: 12570-12574.
6. Zhou Y, Frey TK, Yang JJ (2009) Viral calciomics: interplays between Ca2+ and virus. Cell Calcium 46: 1-17.
7. Herzberg O, Moult J, James MN (1986) A model for the Ca2+-induced conformational transition of troponin C. A trigger for muscle contraction. J Biol Chem 261: 2638-2644.
8. Holmes KC, Popp D, Gebhard W, Kabsch W (1990) Atomic model of the actin filament. Nature 347: 44-49.
9. Mann KG, Nesheim ME, Church WR, Haley P, Krishnaswamy S (1990) Surface-dependent reactions of the vitamin K-dependent enzyme complexes. Blood 76: 1-16.
10. Carafoli E (2002) Calcium signaling: a tale for all seasons. Proc Natl Acad Sci U S A 99: 1115-1122.
11. Sutton RB, Davletov BA, Berghuis AM, Sudhof TC, Sprang SR (1995) Structure of the first C2 domain of synaptotagmin I: a novel Ca2+/phospholipid-binding fold. Cell 80: 929-938.
12. Henikoff S, Greene EA, Pietrokovski S, Bork P, Attwood TK, et al. (1997) Gene families: the taxonomy of protein paralogs and chimeras. Science 278: 609-614.
13. Kawasaki H, Nakayama S, Kretsinger RH (1998) Classification and evolution of *EF*-hand proteins. Biometals 11: 277-295.
14. Grabarek Z (2006) Structural basis for diversity of the *EF*-hand calcium-binding proteins. J Mol Biol 359: 509-525.
15. Bairoch A, Cox JA (1990) *EF*-hand motifs in inositol phospholipid-specific phospholipase C. FEBS Lett 269: 454-456.
16. Finn BE, Forsen S (1995) The evolving model of calmodulin structure, function and activation. Structure 3: 7-11.
17. Nakayama S, Kretsinger RH (1994) Evolution of the *EF*-hand family of proteins. Annu Rev Biophys Biomol Struct 23: 473-507.
18. Gifford JL, Walsh MP, Vogel HJ (2007) Structures and metal-ion-binding properties of the Ca2+-binding helix-loop-helix *EF*-hand motifs. Biochem J 405: 199-221.
19. Godzik A, Sander C (1989) Conservation of residue interactions in a family of Ca-binding proteins. Protein Eng 2: 589-596.
20. Linse S, Brodin P, Johansson C, Thulin E, Grundstrom T, et al. (1988) The role of protein surface charges in ion binding. Nature 335: 651-652.
21. Linse S, Forsen S (1995) Determinants that govern high-affinity calcium binding. Adv Second Messenger Phosphoprotein Res 30: 89-151.

22. Lin HH, Han LY, Zhang HL, Zheng CJ, Xie B, et al. (2006) Prediction of the functional class of metal-binding proteins from sequence derived physicochemical properties by support vector machine approach. BMC Bioinformatics 7 Suppl 5: S13.
23. Zhou Y, Yang W, Kirberger M, Lee HW, Ayalasomayajula G, et al. (2006) Prediction of *EF*-hand calcium-binding proteins and analysis of bacterial *EF*-hand proteins. Proteins 65: 643-655.
24. Franke S, Herfurth J, Hoffmann D (2010) Estimating affinities of calcium ions to proteins. Adv Appl Bioinform Chem 3: 1-6.
25. Boguta G, Stepkowski D, Bierzynski A (1988) Theoretical estimation of the calcium-binding constants for proteins from the troponin C superfamily based on a secondary structure prediction method. II. Applications. J Theor Biol 135: 63-73.
26. Wiseman T, Williston S, Brandts JF, Lin LN (1989) Rapid measurement of binding constants and heats of binding using a new titration calorimeter. Anal Biochem 179: 131-137.
27. Bhattacharya A, Padhan N, Jain R, Bhattacharya S (2006) Calcium-binding proteins of Entamoeba histolytica. Arch Med Res 37: 221-225.
28. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, et al. (2004) The Pfam protein families database. Nucleic Acids Res 32: D138-141.
29. Kunal J, Chandan K, PK N (2010) Prediction of *EF*-hand calcium-binding proteins and identification of calcium-binding regions using machine learning techniques. Journal of Cell and Molecular Biology 8(2): 41-49.
30. VanScyoc WS, Sorensen BR, Rusinova E, Laws WR, Ross JB, et al. (2002) Calcium binding to calmodulin mutants monitored by domain-specific intrinsic phenylalanine and tyrosine fluorescence. Biophys J 83: 2767-2780.
31. Moeschler HJ, Schaer JJ, Cox JA (1980) A thermodynamic analysis of the binding of calcium and magnesium ions to parvalbumin. Eur J Biochem 111: 73-78.
32. Schymkowitz JW, Rousseau F, Martins IC, Ferkinghoff-Borg J, Stricher F, et al. (2005) Prediction of water and metal binding sites and their affinities by using the Fold-X force field. Proc Natl Acad Sci U S A 102: 10147-10152.
33. Wei L, Altman RB (2003) Recognizing complex, asymmetric functional sites in protein structures using a Bayesian scoring function. J Bioinform Comput Biol 1: 119-138.
34. Deng H, Chen G, Yang W, Yang JJ (2006) Predicting calcium-binding sites in proteins - a graph theory and geometry approach. Proteins 64: 34-42.
35. Wang X, Kirberger M, Qiu F, Chen G, Yang JJ (2009) Towards predicting Ca2+-binding sites with different coordination numbers in proteins with atomic resolution. Proteins 75: 787-798.
36. Liu T, Altman RB (2009) Prediction of calcium-binding sites by combining loop-modeling with machine learning. BMC Struct Biol 9: 72.
37. Bock JR, Gough DA (2001) Predicting protein--protein interactions from primary structure. Bioinformatics 17: 455-460.
38. Lo SL, Cai CZ, Chen YZ, Chung MC (2005) *EF*fect of training datasets on support vector machine prediction of protein-protein interactions. Proteomics 5: 876-884.
39. Rout AK, Padhan N, Barnwal RP, Bhattacharya A, Chary KV (2010) Calmodulin-like Protein from Entamoeba histolytica: Solution Structure and Calcium-Binding Properties of a Partially Folded Protein. Biochemistry.

40. Gopal B, Swaminathan CP, Bhattacharya S, Bhattacharya A, Murthy MR, et al. (1997) Thermodynamics of metal ion binding and denaturation of a calcium binding protein from Entamoeba histolytica. Biochemistry 36: 10910-10916.
41. Sigrist CJ, de Castro E, Cerutti L, Cuche BA, Hulo N, et al. (2013) New and continuing developments at PROSITE. Nucleic Acids Res 41: D344-347.
42. Berman H, Henrick K, Nakamura H, Markley JL (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res 35: D301-303.
43. Li W, Jaroszewski L, Godzik A (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics 17: 282-283.
44. Marsden BJ, Shaw GS, Sykes BD (1990) Calcium binding proteins. Elucidating the contributions to calcium affinity from an analysis of species variants and peptide fragments. Biochemistry and Cell Biology 68: 587-601.
45. Gribskov M, McLachlan AD, Eisenberg D (1987) Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci U S A 84: 4355-4358.
46. Henikoff JG, Henikoff S (1996) Using substitution probabilities to improve position-specific scoring matrices. Comput Appl Biosci 12: 135-143.
47. Tatusov RL, Altschul SF, Koonin EV (1994) Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. Proc Natl Acad Sci U S A 91: 12091-12095.
48. Eddy SR (2004) Where did the BLOSUM62 alignment score matrix come from? Nat Biotechnol 22: 1035-1036.
49. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, et al. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci U S A 97: 262-267.
50. Ding CH, Dubchak I (2001) Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 17: 349-358.
51. Chang C-CaL, Chih-Jen (2011) LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2: 27:21--27:27.
52. A. Athanasopoulos AD, V. Mezaris, I. Kompatsiaris (April 2011) GPU Acceleration for Support Vector Machines. Proc 12th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2011).
53. Xiao X, Shao S, Ding Y, Huang Z, Chou KC (2006) Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. Amino Acids 30: 49-54.
54. Xiao X, Wang P, Chou KC (2009) GPCR-CA: A cellular automaton image approach for predicting G-protein-coupled receptor functional classes. J Comput Chem 30: 1414-1423.
55. Bhasin M, Raghava GP (2004) Analysis and prediction of affinity of TAP binding peptides using cascade SVM. Protein Sci 13: 596-607.
56. Ramana J, Gupta D (2010) FaaPred: a SVM-based prediction method for fungal adhesins and adhesin-like proteins. PLoS One 5: e9695.
57. Klein P, Kanehisa M, DeLisi C (1984) Prediction of protein function from sequence properties. Discriminant analysis of a data base. Biochim Biophys Acta 787: 221-226.
58. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157: 105-132.

59. Kuhn LA, Swanson CA, Pique ME, Tainer JA, Getzoff ED (1995) Atomic and residue hydrophilicity in the context of folded protein structures. Proteins 23: 536-547.
60. Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH (1985) Hydrophobicity of amino acid residues in globular proteins. Science 229: 834-838.
61. Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405: 442-451.
62. Boguta G, Stepkowski D, Bierzynski A (1988) Theoretical estimation of the calcium-binding constants for proteins from the troponin C superfamily based on a secondary structure prediction method. I. Estimation procedure. J Theor Biol 135: 41-61.

Chapter 3.

A machine learning approach to modulate the calcium binding affinity in EF-hand proteins and comparative insights into the site-specific binding affinity

3.1 Abstract

Different Ca^{2+}-binding proteins display different levels of affinities for calcium. Many functions of these proteins, in general, depend on their affinity for Ca^{2+}. Earlier, we devised a method to identify and classify the EF-hand loops present in highly conserved and very important class of proteins family. The EF-hand motif is the most frequently occurring Ca^{2+}-binding motif with more than 66 subfamilies plays numerous roles in many crucial pathways. Based on the input sequences, we predicting the loop structure in EF-hand motif and classified the loop into different levels of affinities as high, low or none. The predictor showed high accuracy for the data available from the literature. Utilizing the earlier developed method, we applied the two new scoring schemes to refine the prediction further and designed a EF-hand loop unique is protein database capable of binding calcium with high affinities by mutating residues on the basis of machine learned classifier. The unique sequence was incorporated in the *Entamoeba histolytica* Calcium binding protein1 (*Eh*CaBP1) EF-hand loop 2 using site directed mutagenesis. The sixty-six amino acid residues long protein containing modified second EF-hand loop and the *Eh*CaBP1-Wt with low affinity loops were studied for calcium binding properties using ITC calorimetry. The binding energy indicated at ~535-fold increase in the association constant (K_a) of the designed protein compared to the *Eh*CaBP1-Wt. Furthermore, we used X-ray crystallography to understand the changes at the atomistic level leading to changes in the functional behavior of EF-hand motif in terms of calcium binding. Surprisingly, we found out the high-resolution structure that diffracted at 1.9Å, showed shrinkage in Ca^{2+} binding coordination sphere resulting in strong coordination yielding high affinity for calcium and forming a hexamer due to the structural changes caused by the designed high affinity calcium sequence. We present a set of programs (offline) with new

scoring functions and a user-friendly webserver (online) to predict, design and engineer EF-hand binding loop. The webserver and the downloadable set of scripts are available at http://202.41.10.46/calb/.

3.2 Introduction

Calcium binds specifically at conserved amino acid sites which are designed to bind the divalent ion with very specific binding affinities that defines its function. These amino acid sites are present in various kinds of proteins inside the cell. The calcium dependent regulation is carried out by using calcium as an intracellular secondary messenger by altering the concentration of the ion during cell stimulus [1]. The mechanism that invokes specific role to various proteins depends on the concentration of the Ca^{2+}-ion both in the intra- and extracellular compartments of the cell. Inside the cell, Ca^{2} plays role in the metabolic regulation, muscle contraction, cell motility, nerve transmission, cell division and growth, secretion and membrane permeability[2].

Besides having low sequence identity (33%) between the two proteins; CaM and Nt-*Eh*CaBP1, the structure of both the proteins in complex with Ca^{2+} show two nearly symmetric N- and C-terminal domains connected by a flexible central helix with both the domains contains two EF-loop) hand motifs for binding two calcium ions. The Ca^{2+} binding EF-hand loop site of Nt-EhCaBp1consists of a contiguous stretch of twelve amino acid residues and bind to calcium in pentagonal bipyramidal geometry. The residues in the loop at positions +X, +Y, +Z, -Y, -Z and -X coordinate with calcium, except the 9th (-Z) position whereas water molecule coordinates with Ca^{2+}in many calcium binding proteins. Besides the residues (ligands) in the binding loop being one of the major factor that dictates the calcium binding; other factors such as intrinsic binding affinity of each binding loop, conformational cost upon calcium binding, EF-β-scaffold, cooperative binding, the physiological environment, etc. also play a crucial role in binding of the metal ion. All these factors influence wide range of calcium binding affinities that let these proteins to carry out very specific functions. Orientation of atoms (ligand) in the individual EF loops residues at specific location decide the extent of binding hence influencing the overall binding affinity of the site. The sequence being one of the factors that influences the binding does not

always correlates well with the binding affinity [1, 3, 4]. Hence it is a very difficult task to factor all the dependent variables and predict the exact binding affinity.

Based on the input sequences we have predicted the loop structure in EF-The SVM based approaches using the position specific scoring system have been increasingly popular in not only solving biological problem but to classify a number of different big and complex data problems. Previously [5] we developed a method to identify and classify the calcium binding EF-hand loop. The classifier uses a machine learning support vector machine (SVM) with C and Gamma parameters for a nonlinear SVM with a Gaussian radial basis function kernel [6]. Based on the input sequences, we predicting the loop structure in EF hand motif and classified the loop into different levels of affinities as high, low or none. The predictor showed high accuracy for the data available from the literature [7, 8].

Based on the input sequences, we predicting the loop structure in EF-hand motif and classified the loop into different levels of affinities as high, low or none. The predictor showed high accuracy for the data mined from the literature. A unique sequence was incorporated in the *Entamoeba histolytica* Calcium binding protein1 [9] (*Eh*CaBp1) EF-hand loop 2 using site directed mutagenesis.

Utilizing the earlier developed method, to increase the efficiency of the classification, we extracted the margin distance from the decision boundary for each prediction. Furthermore, we applied these two new scoring schemes to refine the prediction further and designed a EF- hand loop unique to is all the known protein database capable of binding calcium with high affinities by mutating residues on the basis of machine learned classifier scores.

To test the program further we designed a EF-hand loop on the basis of newly devised scoring functions. The unique sequence was incorporated in the *Entamoeba histolytica* Calcium binding protein1 (*Eh*CaBp1) EF-hand loop 2 using site directed mutagenesis. The sixty-six amino acid residues long protein containing modified second EF-hand loop and the *Eh*CaBP1-Wt with low affinity loops were studied for calcium binding properties using ITC calorimetry. The binding energy indicated at ~535-fold increase in the binding of the designed protein compared to the *Eh*CaBP1-Wt. Furthermore, we used X-ray crystallography to understand the changes at the atomistic level leading to changes in the functional behavior of EF hand motif in terms of calcium binding. The high-resolution structure diffracted at 1.9Å showed shrinkage in Ca^{2+} binding coordination sphere resulting in strong coordination yielding high affinity for calcium. We present

a set of programs with new scoring functions and a user-friendly webserver to predict, design and engineer EF-hand binding loop. We validated our findings using biochemical, structural and computational techniques. We present a set of programs (offline) with new scoring functions and a user-friendly webserver (online) to predict, design and engineer EF-hand binding loop. The webserver and the downloadable set of scripts are available at http://202.41.10.46/calb/.

3.3 Materials and Methods

3.3.1 SVM Algorithm

Previously, we developed a method to identify and classify the calcium binding EF-hand loop. The classifier uses a machine learning support vector machine (SVM) with C and Gamma parameters for a nonlinear-SVM with a Gaussian radial basis function (RBF) kernel. In order to increase the efficiency of the classification, we extracted the margin distance from the decision boundary for each prediction. Margin here is referred to the distance of the vector (Ca^{2+} -binder /non-binder /high affinity/ low affinity binder) x, which is the Euclidean distance of x from the separating hyperplane, is given by:

$$\frac{w.x+b}{norm(w)}$$

Where, w = perpendicular to the hyperplane and b is the parameter to determines the offset of the hyperplane from the origin along the normal vector.

The distance to the origin is calculated using $\frac{|b|}{||w||}$. The margin could be defined as the summation of$d(margin) = m_positive + m_negative$. The *m_positive* is the distance from the hyperplane to the high Ca^{2+} binding site and the *n_negative* is the distance from the hyperplane to the low binding in the binding affinity prediction program. The distances calculated for each site prints, can be either positive or negative depending on the classification. In general, the distance shows how far or close the binding sites are have been classified by the kernel function [10, 11].

3.3.2 Log-Odds Substitution Scores

The conservation score based on PSSM for each sequence submitted for classification were calculated by$Gsij = \log\ (Sij/Pi)$. Bij is the probability of amino acid i at position j in matrix S, calculated by using the equation

$$Bij = q + \frac{bPi}{n} + b$$

Where q = observed counts of amino acid type i at position j, Pi = probability of amino acid type i, b = pseudo count which is considered here as square root of the total number of training sequences and n is the number of training sequences. The calculated *Gsij* (PSSM Score) represents the foreground model which representing true homology and Pi represent chance that a match occurs at random (background model) calculated using BLOSUM62 substitution matrix[12]. In concise, the PSSM based scoring includes the relative frequencies obtained by counting the occurrence of each amino acid at each position of the alignment, followed by normalization of the frequencies[5].

3.3.3 Designing of unique EF-loop site

We used SVM margin scores (SVM_{Mar}) and the PSSM based log likelihood scores (PSM_{LogL}) to design a unique calcium binding site which is not present in any of the protein sequence databases. The iterative database search upon each computation point mutation was performed by BLASTp [13]. We designed the same mutant (DKDGDGFIDFEE) by using SDM in second EF-hand loop of Nt*Eh*CaBP1. The Nt*Eh*CaBP is a EF-hand containing calcium binding protein which is well characterized in our lab [9, 14] and has two calcium binding sites. In order to construct the desired mutant, we incorporated five point mutations in the second EF-loop (DADGNGEIDQNE) of EhCaBP1-Nt. The mutations were incorporated at the following positions: A47K, N50D, E52F, Q55F and N56E. We selected Nt*Eh*CaBP1 and the Nt*Eh*CaBP1 *EF*2 site as the model in this study based on the following criteria; 1) the sequence similarity with the 2nd EF-loop. 2) Availability of biophysical and structural data. 3) The Nt*Eh*CaBP1construct has two calcium binding sites which allowed us to observe the cooperative binding in EF-hand motifs in two protein constructs with

one with loop constructed for higher binding affinity (Nt*Eh*CaBP1 EF2) compared to low binding sites in Nt*Eh*CaBP1. The design aiming with an idea of improving the classifiers predictions, understand cooperative binding and to improve the false positive predictions; we chose Nt*Eh*CaBP1 EF2 loop for experimental validation.

3.3.4 Cloning of Nt EhCaBP1EF-2 mutant

The gene fragments corresponding to the N-terminal domains of *Eh*CaBP1 protein were cloned by using existing N-terminal clone of *Eh*CaBP1 as a template in the bacterial expression vector, pET 28(b). The mutations were created (in *Eh*CaBP1 *EF*-II) at site 141, 150, 156, 165, 168 by site directed mutagenesis. The following primers were used for the mutation.

CaBP1 Mut K2 FP 5'-CAAATCTATTGATAAAGATGGAAATGG-3'
CaBP1 Mut K2 RP5'-CCATTTCCATCTTTATCAATAGATTTG-3'
CaBP1MutD5, F7 FP 5'CTATTGATAAAGATGGAGATGGATTTATTGATCAAAATGAATTTGC-3'
CaBP1MutD5, F7 RP 5'-GCAAATTCATTTTGATCAATAAATCCATCTCCATCTTTATCAATAG-3'
CaBP1Mut F_{10} FP 5'- ATTTATTGATTTTAATGAATTTGC-3'
CaBP1Mut F_{10} RP 5'-GCAAATTCATTAAAATCAATAAAT-3'
CaBP1Mut E_{11} FP 5'-ATTTATTGATTTTGAAGAATTTGC-3'
CaBP1Mut E_{11} RP 3'-GCAAATTCTTCAAAATCAATAAAT-3'
CaBP1 FP -5'-CATGCCATGGCAATGGCTGAAGCACTTTTTAAAG-3'
CaBP1 RP-5'-CGGCTCGAGGAGTGAAAACTCAAGGAATTCTTC-3'

After mutating all five residues, insert was cloned in pET-28b vector. Mutations were confirmed by sequencing.

3.3.5 Overexpression and Purification of NtEhCaBP1 EF-2 mutant

The NtEhCaBP1 was expressed and purified as described in previously published literature. The NtEhCaBP1 *EF*-2 mutant construct was transformed into *E.coli* strain BL21 (DE3) for expression. Cells were grown in LB medium supplemented with 50 mg ml^{-1} kanamycin at 37˚C. The culture was induced with 0.8mM IPTG, when the OD reached 0.7 at A_{600}. It was then incubated at the same temperature for 3h for further growth. Cells were harvested by centrifugation at 7000 rpm for 10 min. The cell pellet was resuspended in suspension buffer (50mM Tris pH 7.5, 2mM

EGTA). The cells were then lysed by freeze-thaw followed by sonication. Clear supernatant was obtained by centrifugation at 12000 rpm for 30 min. The protein supernatant was passed through ion-exchange chromatography column pre-equilibrated with ten bed volumes of suspension buffer for ion-exchange purification. It was then washed with 30-40 ml wash buffer (50mM Tris pH 7.5, 5mM NaCl,) to remove non-specifically bounded proteins. Finally, the protein was eluted with elution buffer (50mM Tris pH 7.5, 5mM $CaCl_2$). Further, ion-exchanged purified protein was subjected to gel filtration chromatography in buffer containing 50mM Tris pH 7.5, 5mM $CaCl_2$. The purity of the protein was checked using 15% SDS-PAGE. The purified and dialyzed protein was concentrated to 15 mg ml^{-1} using a 3 KDa cut-off centricon prior to crystallization.

3.3.6 Preparation of Ca^{2+} free of NtEhCaBP1 EF-2 Mutant and Native NtEhCaBP1

Both proteins were prepared in Ca^{2+} free (apo form) as described before (Shivesh *et al*., 2012) and the proteins were dialyzed against 10mM of HEPES pH 7.4.

3.3.7 Isothermal titration Calorimetry (ITC) to calculate dissociation constant of Calcium

ITC experiments were carried out using Microcal ITC $_{200}$ instrument from GE-health care. Experiments were performed at 30˚C in 10mM HEPES buffer (pH 7.4). Same buffer was used in the reference cell. All solutions were thoroughly degassed by stirring under vacuum before use. 1mM $CaCl_2$ was titrated into 3.4µM of Nt*Eh*CaBP1 and Nt*Eh*CaBP1 *EF*-2 mutant, in the 280µl of sample cell. For each titration, CaCl2 in 10mM HEPES and protein in 10mM HEPES were used as titrant and analyte (known concentration) respectively. The volume of each titrant sample was 2ul per injection and because of fixed titrant pipette volume (40µl) led us to perform 20 injections of titrant to sample cell containing analyte (known concentration) as protein. The mixture was allowed to react for 2 min between injections. Heat release due to the injection and dilution were obtained by titrating $CaCl_2$ into buffer containing protein. Raw thermogram was generated with rate of change of heat with respect to time by software Microcal iTC200. The data were fitted using the modified Origin software Microcal Analysis Launcher supplied by Microcal. The plot of heat change with respect to molar ratio of [ligand] / [protein] was derived from raw thermogram

by Origin software Microcal Analysis Launcher. Iterations were performed till the chi square value became reduced and stable. The legend of thermogram with model fitting as one set of sites suggested the reaction of Ca^{+2} to EhCaBP1(both (Both native and mutated) as a sequential binding mode having N (Number of binding sites) as two. ΔH (enthalpy change), ΔS (entropy change) and K_a (association constant) values were obtained as legend by fitting the theromgram with reference to a binding isotherm model fitting as sequential binding sites (N=2) gave a nonlinear curve trend line with the dataset.

3.3.8 Crystallization of NtEhCaBP1 EF-*2 Mutant*

Crystallization was carried out by the hanging drop vapour diffusion method in 24 well plates using 2µl of protein solution was mixed with an equal volume of precipitant solution and equilibrated against 500 ml reservoir solution (precipitant). Initially the same crystallization condition was used in which native Nt*Eh*CaBP1 (N-terminal EhCaBP1) was crystallized[9]. We could not get crystals in native Nt*Eh*CaBP1 crystallization condition, rather the condition was closer to native *Eh*CaBP1 crystallization condition [9](Kumar *et al*., 2007). The Nt*Eh*CaBP1 *EF*-2 Mutant was crystallized in MPD 58%- 63% sodium actetate buffer pH 5.0-5.5 with 5mM $CaCl_2$.

3.3.9 X-ray diffraction, Data Collection, processing and structure solution

Crystals were soaked in cryo-protectant solution consisting of 65% MPD, 100mM sodium acetate pH 5.3, 5 mM$CaCl_2$. Single crystals were picked up in cryo loops and flash-cooled in liquid nitrogen. Higher resolution data was collected at ESRF DBT-BM14 France. The crystals diffracted to 1.9Å resolution. Diffraction data were processed and scaled using HKL2000[15]. The crystals belonged to space group $P2_12_12_1$, with unit cell parameters a= 44.6, b= 101.3, c= 107.4 Å. The Matthews coefficient, V_M was 2.90 $Å^3Da^{-1}$, indicating the presence of six molecules in the asymmetric unit, with a solvent content of 57.5%. The structure was solved by molecular replacement with Phaser program [16] using the native structure of EhCaBP1 (2NXQ) as the search model and assembled trimer was used for molecular replacement, the structure solution resulted six molecules in asymmetric unit. Twelve calcium atoms, (two calcium ions in each chain) were identified in the electron density in the center of the EF-hand loop and included in the

refinement. The structure was refined by iterative model building using COOT graphics package combined [17] with Translation, Liberation and Screw-rotation (TLS) displacement parameters restrained refinement was performed. For the final model, the R_{work} was 22.1 % and R_{free} was 26.6%. The structure had good stereochemistry as indicated by program PROCHECK [18] with 97.6% of residues lying in the most favored regions of the Ramachandran plot. The data collection and final refinement statistics are shown in **Table 3.1**.

DATA SET	**Nt*Eh*CaBP1 (Published)**	**Nt*Eh*CaBP1-EF2 mutant**
Crystallographic data		
X-Ray Source	Microstar	ESRF BM14
Wavelength (Å)	1.5418	0.97
Space group	P_3	$P2_1\ 2_1\ 2_1$
Unit Cell Parameters		
a, b, c (Å)	89.589, 89.589, 35.049	44.69, 101.36, 107.47
α, β, γ (º)		90, 90, 90
Resolution (Å)	2.5	1.90
Resolution range (Å)	25 – 2.5	73.44 – 1.90
Completeness (%)	99.0(99.5)	99.86
R_{merge}		
	10.3(1.6)	1.95(at 1.89 A)
Total No. of observations	39681	301767
No. of unique observations	10746	37405
Redundancy	3.68	8.06
Refinement Statistics		
R factor (%)	23.3(23.9)	21.0
Free_R (%)	27.1(27.5)	25.0
B factor	51.0	31.2

3.3.10 Structure and sequence analysis

We performed the sequence alignments using Clustal Omega [19] and BioEdit [20] programs. The structural alignment was performed using the Dali server [21]. Protein – protein interactions were calculated using the *PDBsum* webserver[22]. The calcium coordination distances and angles were calculated using LIGPLOT [23] and PyMol [24, 25] software. The images were prepared in PyMol, Chimera and Photoshop software's [24].

3.4 Results

3.4.1 Algorithm

PSSM based methods are widely used in machine learning based approaches and are well established in classifying data[26, 27]. In the earlier study involving the prediction of binding affinity, we extracted many EF- hand loop sequences along with their binding affinities (K_a). In this study, we implemented the SVM_{Mar} and the PSM_{LogL} scores for all the sequences of the three different datasets consisting of 131 unique Ca^{2+} binding EF- hand loops and ranked them on the basis of the literature reviewed. The details of the predictions for each site along with the scores are shown in the Appendix I.

3.4.2 Designing the high binding affinity EF-hand loop

We incorporated two scoring schemes to categorize the EF- hand loops on the basis of their binding affinities. These scores were computed using the SVM margin function and PSSM based log likelihood algorithm. In order to validate our prediction methods by utilizing the newly devised scoring scheme, we attempt to manipulate the binding affinity of a known calcium binding site to validate our predictions. Therefore, we designed a unique EF-loop, not present in any protein sequence database (using iterative BLASTp search upon each mutation). The designed binding

loop with amino acids 1-DKDGDGFIDFEE-12 showed a SVM score of 2.694 and PSSM score of 6.46.

Our lab works on the biophysical and structural studies of calcium binding proteins (CaBPs) from *Entamoeba histolytica* (EHI-IMSS strain). Out of the 28-predicted calcium binding EF-hand proteins in *EHI[14]*, we have successfully biophysically characterized *Eh*CaBP1, 3 and 5 [28, 29]. To insert the EF-loop sequence that has the predicted high calcium binding affinity, we mutated the sequence of the 2nd EF-hand motifs' calcium binding loop from *Eh*CaBp1. The *Eh*CaBP1-Nt-EF2 loop was selected on the basis of sequence similarity with the designed loop compared to other characterized proteins in *E. Histolytica*. The crystal structure of the N-terminal construct of Nt*Eh*CaBP1 has two Ca^{2+} binding sites and the full-length protein has four calcium binding EF-hand motifs[9]. ITC experiments on the full-length protein suggested four binding sites with two sites having high Ca^{2+} binding affinity and two sites with low binding affinity[30]. The first calcium binding motif that is the EF-1; (1st EF-loop-DVNGDGAVSYEE) has a SVM_{Mar} and PSM_{LogL} score of *-1.045* & *4.976* is predicted to have lower binding affinity. The 2nd EF-loop DADGNGEIDQNE with SVM_{Mar} and PSM_{LogL} scores of 1.001 is predicted to have relatively high binding affinity compared to the EF-1 loop. We considered designing a construct with two EF-hand loops to understand the mechanistic details of cooperative binding, a phenomenon that enables a pair of EF-hand to bind Ca^{2+} with high binding affinity compared to the binding with one EF-hand. In order to figure out further about the characteristics of the designed loop we purified the *Eh*CaBP1-Nt-*EF*2 mutant protein.

3.4.3 The solution state of the mutant suggests it is an oligomer

The recombinant protein was purified by lysis through freeze-thaw cycles and followed by sonication and collection of pure supernatant via centrifugation. The crude supernatant was used in order to purify Nt*Eh*CaBP1-EF2 mutant by Ion-Exchange chromatography. Ion-Exchange purified protein was then subjected to a second purification process i.e. the size exclusion chromatography (Gel Filtration chromatography) using a buffer containing 50mM Tris-Cl (pH 7.5) and 5mM $CaCl_2$ employing supertax G-75 column of FPLC supplied by GE. The Nt*Eh*CaBP1-EF2 mutant protein peak was noted at 65.93ml (Figure 3.1).

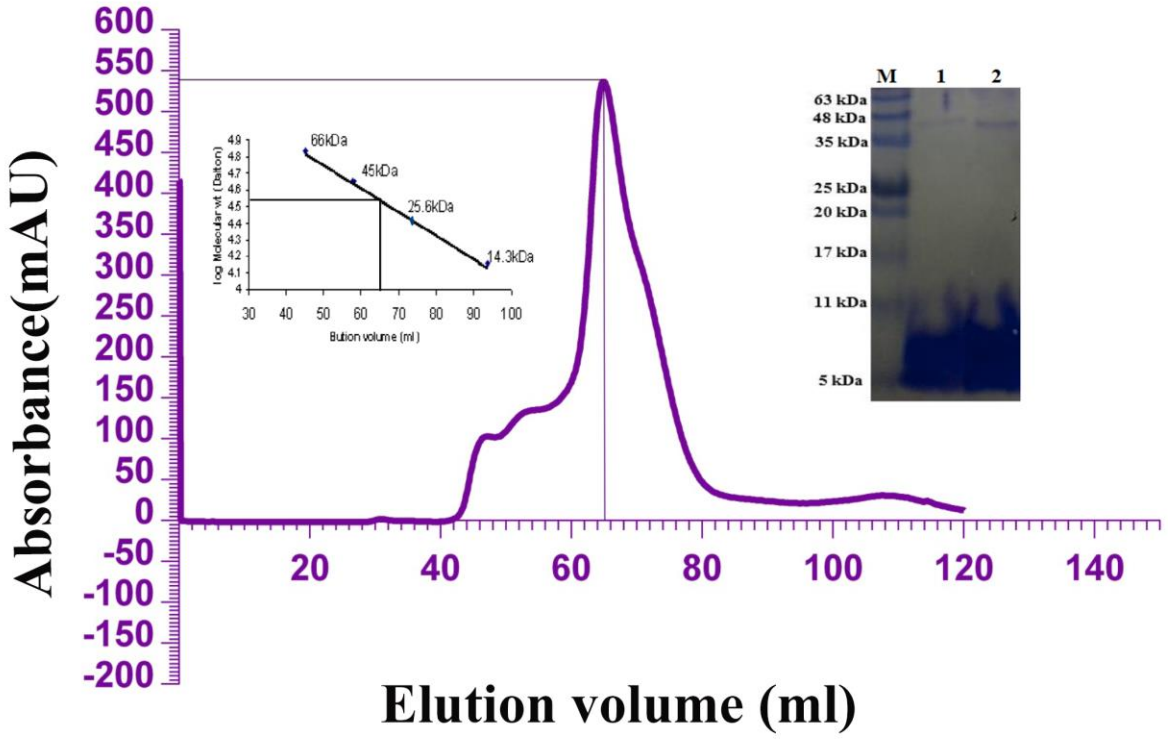


Figure 3.1 FPLC profile of NTD-*Eh*CaBP1 EF2 mutant. The peak of NTD-*Eh*CaBP1 EF2 mutant on sephadex-75 column was eluted at 64.931mL volume corresponding to 39KDa molecular weight. The Weight of Mut-NtdCabp1 calculated on the basis on the amino acid sequences was 7.23KDa.

The standard curve verses void volume were plotted and calculated for the same G-75 column prior to experiment using standard proteins supplied by Sigma Aldrich of different molecular weights. The elution fraction was run to check protein purity. The protein was overexpressed in *E. Coli* BL21 (DE3) and the expressed protein was purified by using Ni-NTA chromatography followed by gel filtration chromatography. The gel filtration profile clearly indicated hexameric nature of the protein with a molecular mass of about 42 KDa. SDS-PAGE indicated the expected molecular mass of 7.0 KDa for one monomer (Figure 3.1).

3.4.4 ITC Isotherms shows clear distinction in calcium binding pattern

To check whether the mutation enhances the Ca^{2+} binding affinity in EF-2 loop of Nt-*Eh*CaBP1 mutant compared to the Nt-*Eh*CaBP1, we performed ITC experiment. The LogK values were obtained by fitting a macroscopic binding model using sequential binding mode with two binding sites as observed.

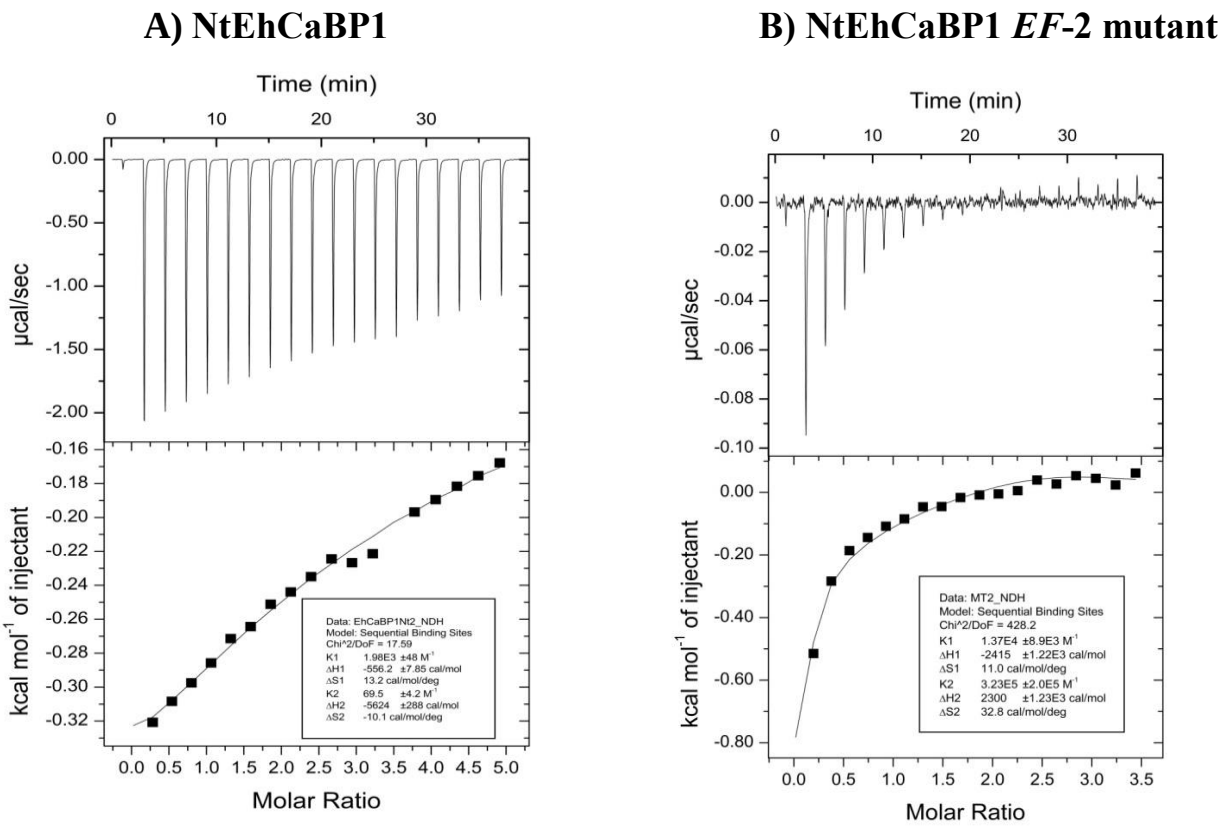


Figure 3.2 Isothermal titration calorimetric analysis of Ca^{2+}-binding to the holo-form of Nt-*Eh*CaBP1 and plot of kcal mol^{-1} of heat absorbed/released per injection of $CaCl_2$ as a function of molar ratio of Ca^{2+}: protein at 25°C is shown. For all titrations, the top panels represent the raw data (power: time) and the bottom panels represent integrated binding isotherms. The solid line represents the best nonlinear fit to the experimental data. Binding isotherm for A: Nt*Eh*CaBP1; B. Nt*Eh*CaBP1-EF2 mutant. Thermodynamic parameters obtained are summarized in the Table 3.2.

The concentration of apo-proteins was determined by nanodrop (Thermo Nanodrop 2000) exploiting molar extinction coefficient and molecular weight calculated by Protparam [31]. The stock solution of protein and ligand were prepared prior to titration in the degassed buffer. Isothermal titration was performed in buffer to check the background thermal noise of the buffer dilution. In order to decrease the thermal noise, reference cell was filled with the same buffer and both ligand and protein were prepared in the same buffer. The heat released and absorbed during titration was calculated by the equation:

$$\mathbf{q = V\Delta H\Delta [L_B]}$$

q= Heat change according to change in ligand bound conc.

$\Delta[L_B]$= Change in bound ligand conc.

ΔH= change in enthalpy after ligand binding

V= reaction volume

The representative curves (Figure 3.1) and the derived binding isotherms for the binding of Ca^{2+} to apo-proteins were made to evaluate the binding association, enthalpy and entropy changes related to the reactions. The titrations stoichiometry of reaction suggested a 2:1 protein ligand ratio. The isotherm plot of the calorimetry experiments showed a clear indication that the binding of calcium was different in both the proteins. The titration repeated three times (data not shown) for each system showed a similar trend. The trend line suggests that Nt*Eh*CaBP1 EF2 mutant over the time saturated well and reached the baseline compared to the native Nt*Eh*CaBP1 which did not saturate properly (Figure 3.2).

The K_a (association constant) calculated after the titrations are listed in Table 3.2 shows an increase of ~160 fold in calcium binding if we compare the constants of EF-loop site1 and a massive increase of ~535 folds in the second site of EF-loop2. The association constant for Nt*Eh*CaBP1 for the site L1= (0.695± .042) $*10^2$ and L2= (1.98± 0.048) compared to the Nt*Eh*CaBP1 EF2 mutant where site H1= (3.72± 0.0008) $*10^4$ and H2= (3.23± 2.0) $*10^5$ indicates the massive change in the binding affinity of the EF2 mutant that five different residues at the 2nd EF-2 loop.

Protein-Ligand	**Number of experimental sites(n)=2**	**Ka(M^{-1})**	**ΔH (cal/mol)**	**ΔG (Cal/mol)**	**ΔS (cal/ mol/ deg)**
NtEhCaBP1	Site-1	L1= (0.695± .042) $*10^2$	-5624 ± 288	-5321± 288	-10.1
	Site-2	L2= (1.98± 0.048) $*10^3$	-556.2 ± 7.85	-3522 ± 7.85	13.2
NtEhCaBP1-*EF*2-mutant	Site-1	H1= (3.72± 0.0008) $*10^4$	-2415 ± 1.22E3	-2745± 1.22E3	11.0
	Site-2	H2= (3.23± 2.0) $*10^5$	2300 ± 1.23E3	-1316± 1.23E3	32.8

Table 3.2. Summary of macroscopic binding constants and thermodynamic parameters obtained from the ITC studies of Ca^{2+}-binding isotherm of Nt*Eh*CaBP1 and Nt-*Eh*CaBP1 EF-2 mutant at 25°C.

The binding of Ca^{2+} to the Nt*Eh*CaBP1 appeared to be an exothermic process with favorable enthalpy in both the sites (Δ*H*, -556.2 and -5624 kcal/mol). The overall change in entropy in site1 was unfavorable (Δ*S*, ~ −10 cal mol^{-1} K^{-1}) and compared to site2 that shows favorable entropy (Δ*S*, ~ 13.2 cal mol^{-1} K^{-1}). In the Nt*Eh*CaBP1-*EF*2-mutant the enthalpy recorded for the site 1(Δ*H*, -2415 and 2300 kcal/mol) showed that it is an exothermic reaction with favorable enthalpy and site 2 showed (2300 kcal/mol) unfavorable enthalpy indicating that the binding is an endothermic process. Both the sites in the EF-2 mutant showed favorable entropy of Δ*S*, ~ 11.0 and 32.8 cal mol^{-1} K^{-1}.

3.4.5 Crystal structure of NtEhCaBP1 EF-II mutant has six molecules asymmetric unit

To obtain a better insight into the calcium binding mechanism, we carried out crystallization trails on the mutant construct. The structure of the native is already published. Surprisingly, the mutant protein did not crystallize in the same condition which is used for the crystallization of Nt*Eh*CaBP1. The N-terminal *Eh*CaBP1 mutant protein crystallized in MPD 58%- 63% sodium acetate buffer pH 5.0-5.5 with 5mM $CaCl_2$. The structure was solved by molecular replacement with Phaser program [16] by Nt*Eh*CaBP1[9] (PDB 2NXQ) as the search model. The final model is refined up to R_{work} 22.1% and R_{free} is 26.6%. The model is refined with good electron density.

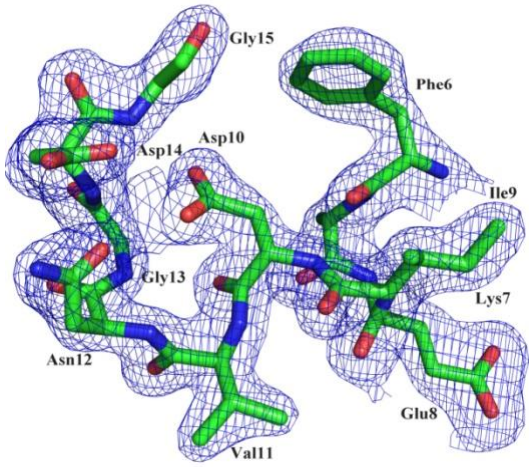


Figure 3.3 Image showing the electron density map of few residues at 1.5σ cutoff, in which Phe6, Lys7, Glu8, Ile9, Asp10, Val11, Asn12, Gly13, Asp14and Gly15 are represented in the electron density.

The final refined and deposited (PDB: 5XOP) model shows that the two EF-hand motifs of Nt*Eh*CaBP1 *EF*-2 mutant were separated by a long helix as seen in many calcium binding proteins.

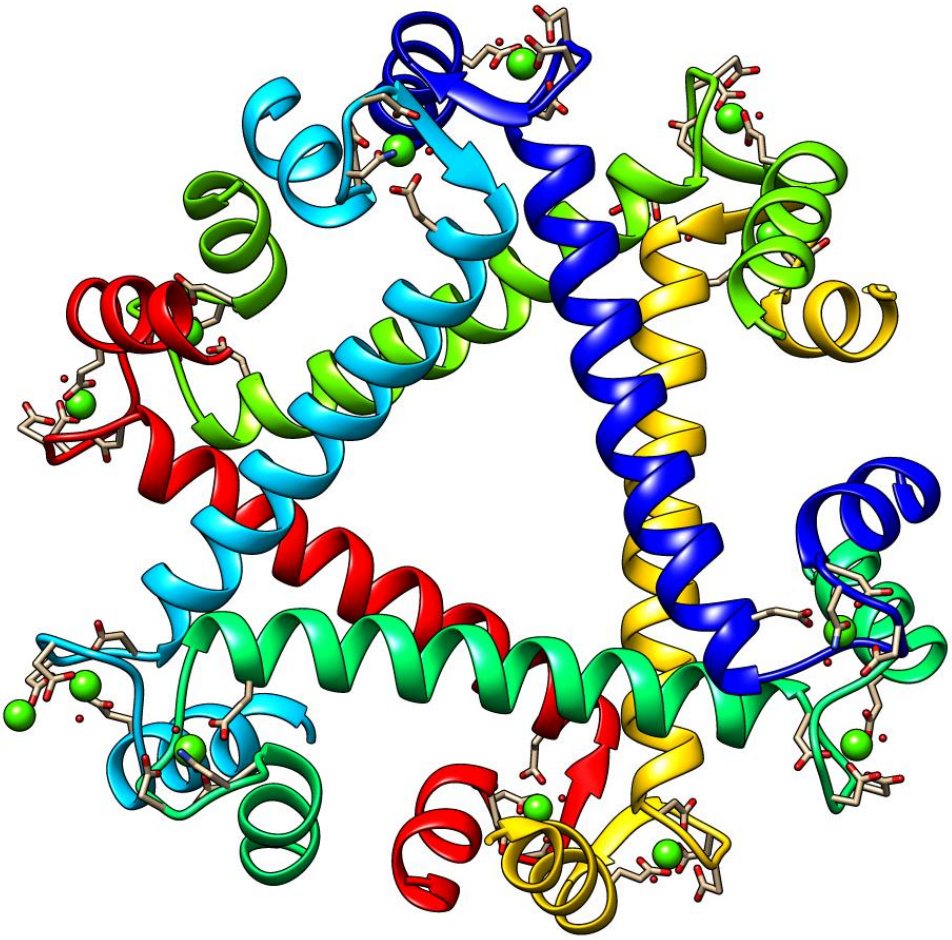


Figure 3.4 The crystal structure of Nt*Eh*CaBP1 EF-II mutant. The structure of Nt*Eh*CaBP1 EF-II mutant shows six molecules of protein forming a hexamer. Negatively charged residues interacting with Ca^{2+} (green) are shown in stick representation. Two EF-hand motifs of Nt*Eh*CaBP1 EF-2 mutant were separated by a long helix, similar to that of crystal structure of N-terminal domain of Nt*Eh*CaBP1 where one Ca^{2+} is bound at each EF-hand motif except the 2nd subunit (chain B) where 2 Ca^{2+} were observed.

3.4.6 Comparison with the Native NtEhCaBP1 shows a bend in the third helix

The Nt*Eh*CaBP1 EF-2 Mutant structure contains six molecules in asymmetric unit, where two trimers interact with each other and form a stable hexamer as seen in the crystal structure (Figure 3.4) unlike the native N-terminal *Eh*CaBP1 structure which forms trimer [9]. The individual chains

of both the proteins were superposed by align () command in PyMol [25]. The superimposition showed the alignment of all the 66 cα atoms yielding a RMSD of 0.96Å (Figure 3.5A). The difference in the orientation of the third α-helix was evident from the alignment. The same helix was closer to the mutated calcium binding site. We calculated the change in orientation by taking a reference point from the central helix and the second reference point was taken from the perpendicular helix (Figure 3.5B) followed by the last residue of the C-terminal end. The change in the angle taken from the same reference suggested that the C-terminal (3rd α-helix) moved around ~7 degrees away from the NtEhCaBP1.

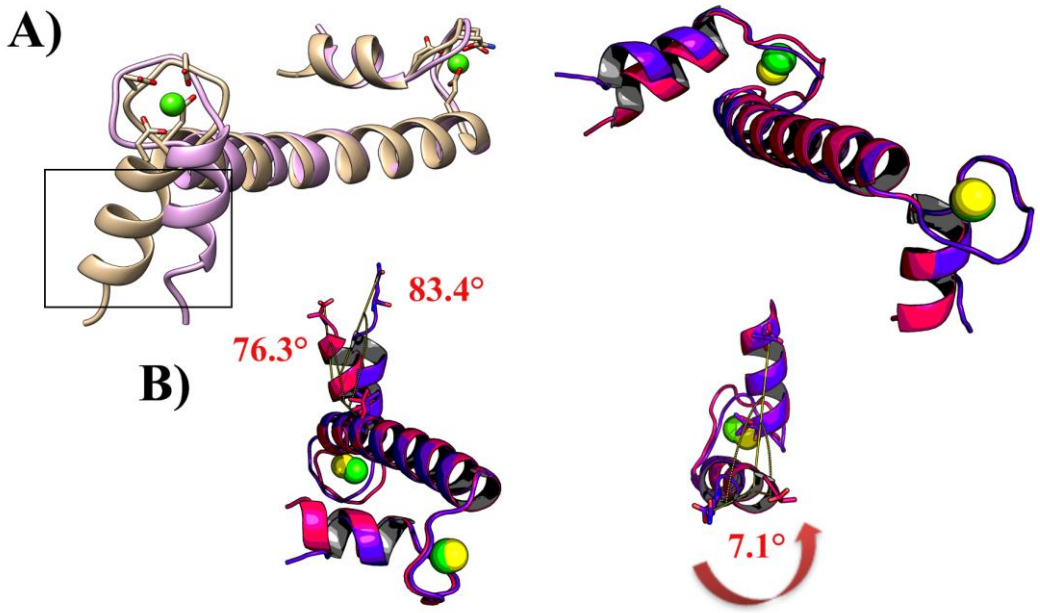


Figure 3.5 The superimposition of Nt*Eh*CaBP1 and Nt*Eh*CaBP1 EF-2. A) The structure of the Nt*Eh*CaBP1 is shown in golden color and Nt*Eh*CaBP1 EF-2 shown in pink color. The figure shows two different orientations of the same alignment. B) The differences in the orientation of the c-terminal helix are calculated with reference to the central helix. The change of the orientation of the c terminal helix is shown in two different orientations where Nt*Eh*CaBP1 is shown in violet color and Nt*Eh*CaBP1 EF-2 shown in pink color.

3.4.7 Calcium induced oligomerization in NtEhCaBP1-EF2 mutant

The N-terminal structure of the native Nt*Eh*CaBP1 revealed a trimeric arrangement (Figure 3.7A) with molecules interacting in a head-to-tail manner (Figure 3.7) forming an assembled domain at the interface with EF1 and EF2 motifs. The full-length structure of native Nt*Eh*CaBP1 is still not crystallized probably due to the presence of highly disordered regions in the C-terminal end however the N-terminal domain can carry out most of the functions of full-length protein [9].

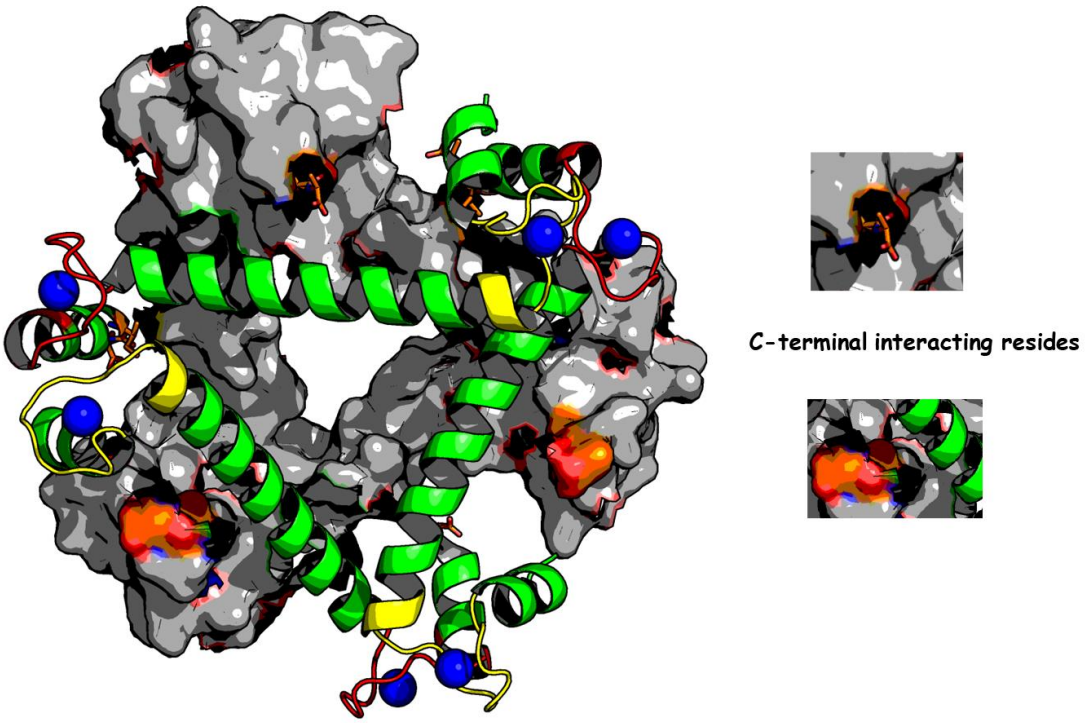


Figure 3.6 The binding interface-I of trimer 1 (Chain A, B, C) and 2 (D, E, F) forming hexamer (Nt*Eh*CaBP1 EF-2). The subset in the image shows the binding hotspots in the C-terminal of every subunit in close proximity with residues central helix residues.

The extended conformation of Nt*Eh*CaBP1 EF-II mutant formed a domain swapped trimer exactly similar to native N-terminal domain structure, where three symmetry-related molecules interacted in a head-to-tail manner which lead to trimerization of N-terminal domain. Surprisingly, in the case of the Nt*Eh*CaBP1 EF-II mutant, due to the bend in the helix-III, one trimer (interface-A) gets close to the other trimer (interface-B) and forms hexamer (Figure 3.6).

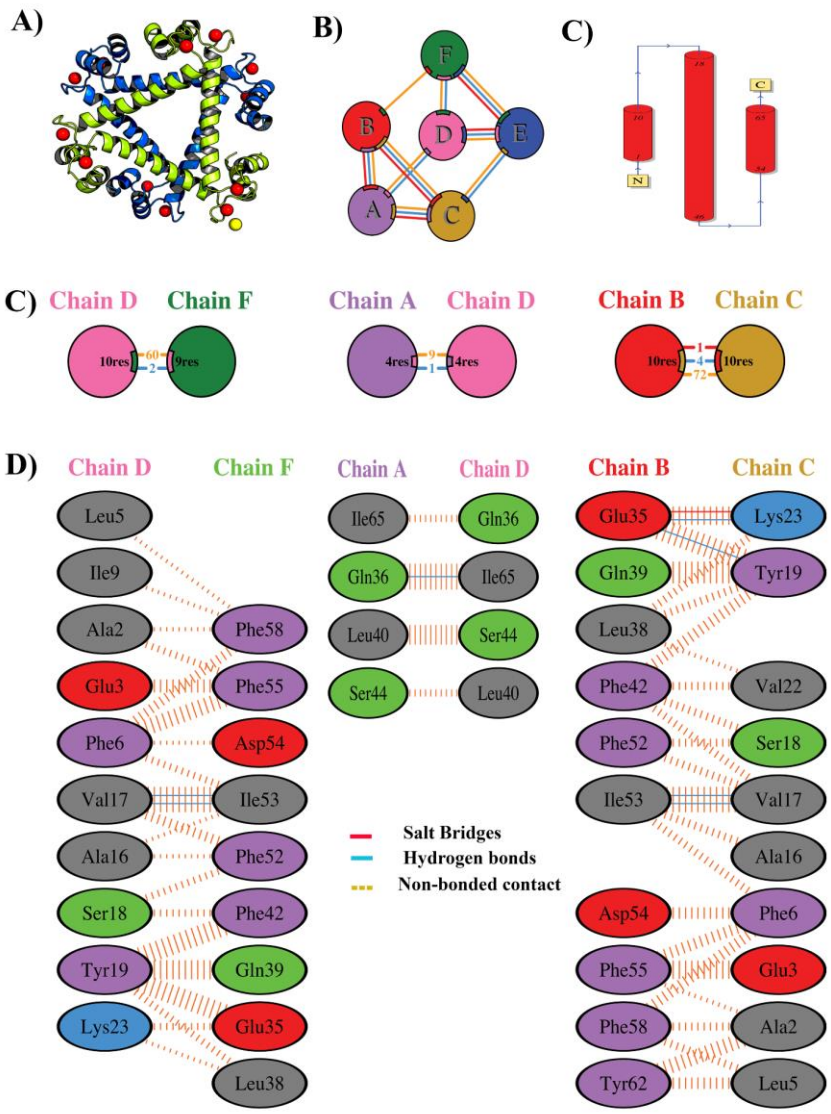


Figure 3.7 The hexametric assembly and the interactions of each of the subunits. A) The crystal structure showing the hexamer of Nt*Eh*CaBP1 EF-II mutant. B) The contacts shared by all the six subunits are represented. C) The secondary structure and molecular topology of the hexamer. D) The hexamer is represented with the chain identifier and the protein – protein interactions are shown with the interaction.

Table 3.3. The interaction of all the individual chains with the binding interface (area), number of residues participating, type of possible interactions is listed for the hexameric complex of NtEhCaBp1-*EF*2 mutant. The statistics is shown for both the trimeric-trimeric (3:3) interface as well as the monomeric interface (1:1).

Chains	No. of interface residues			Interface area (Å^2)			Salt bridges	Hydrogen bonds	Non-bonded contacts
A B	10	:	11	597	:	566	1	4	75
B C	10	:	10	587	:	573	1	4	72
D E	12	:	10	618	:	628	1	4	74
A C	11	:	10	562	:	578	1	4	64
A D	4	:	4	286	:	262	-	1	9
B F	6	:	5	299	:	303	-	1	9
C E	5	:	5	315	:	298	-	1	17
D F	10	:	9	545	:	569	-	2	60
E F	10	:	12	588	:	568	1	4	78

The two assembled domains each trimer interacts with the other trimer and forms star like arrangement of two trimers (Figure 3.7A). The two trimers show a large number of non-bonded contacts (long range interactions) possibly due to the unique arraignment of the trimers (Figure 3.7B). The Chain A interface interacts with Chain D (Figure 3.7). In the same manner chain B interact to chain F and chain C interact to chain E, all the interacting residues are same as they interact in case of chain A and D. These interacting residues forms hydrogen bonds and weak interactions are also observed within the interacting residues. The critical residues which are involved in hydrogen bond interactions (Table 3.4) are the residues from the terminal end of the

subunits which interacts with glutamate-36 from the central helix. All the interactions between all the subunits between interface –I and Interface –II are shown in Figure 3.7C and D.

Acceptor		Donor		Distances
GLN-36 (NE2)	Chain A	Ile-65 (O)	Chain D	3.06
GLN-36 (NE2)	Chain B	Ser-64 (O)	Chain F	3.36
GLN-36 (NE2)	Chain C	Ile-65 (O)	Chain E	3.28

Table 3.4 The hydrogen bonds formed between the Trimeric Interface-I (Chain: A, B, C) and Trimeric Interface-II (Chain: D, E, F).

The interactions of all the three subunits in the trimeric structure of *Eh*CaBP1 are quite similar to that of the hexameric mutant. The interface areas as well as the residues participating in binding are similar with only difference is the occurrence of one additional hydrogen bond amongst the chains in trimers of hexamer (Table 3.3 & 3.5). The list of all the crucial residues for the hexameric and trimeric assembly are graphical shown in Figure 3.7 and 3.8.

Chains	**No. of interface residues**			**Interface area (Å^2)**			**Salt bridges**	**Hydrogen bonds**	**Non-bonded contacts**
A B	**10**	**:**	**10**	**565**	**:**	**572**	**1**	**3**	**74**
B C	**10**	**:**	**10**	**565**	**:**	**572**	**1**	**3**	**74**
A C	**10**	**:**	**10**	**572**	**:**	**565**	**1**	**3**	**74**

Table 3.5. The interaction of all the individual chains with the binding interface (area), number of residues participating, type of possible interactions is listed for the trimeric complex of Nt*Eh*CaBP1.

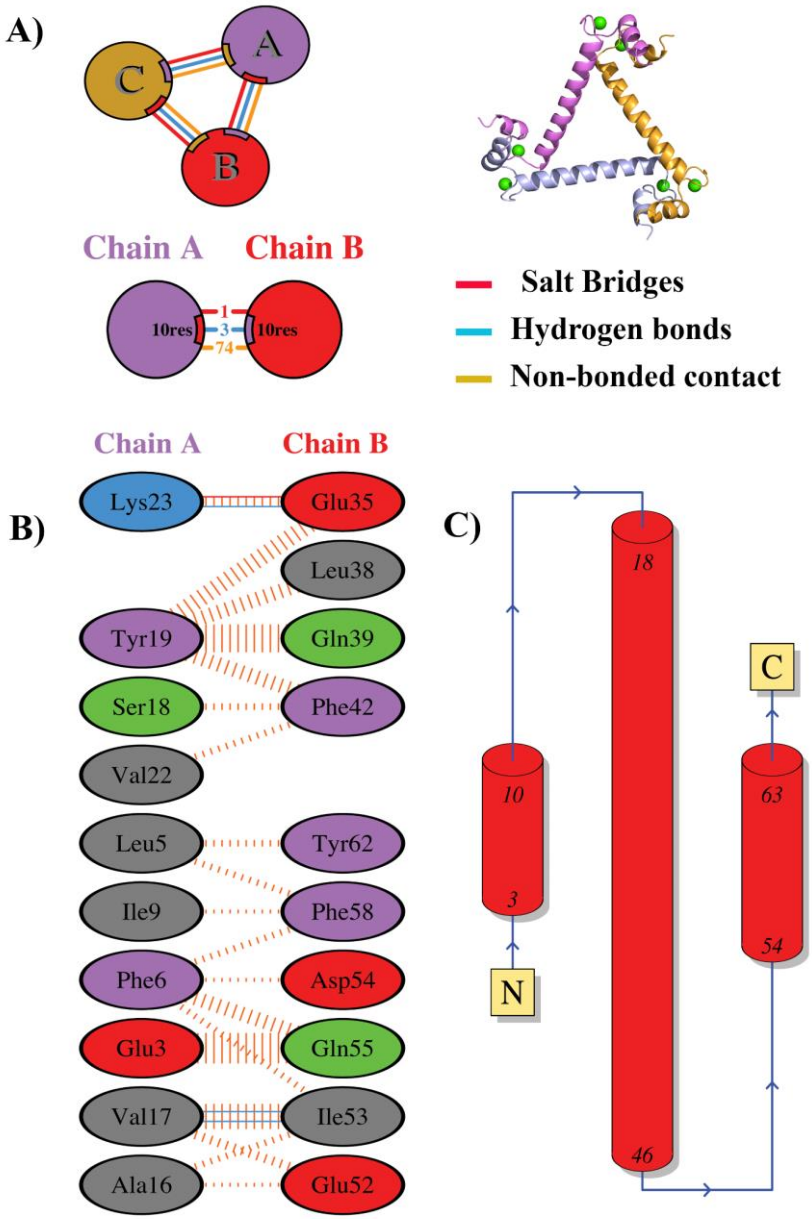


Figure 3.8 The interaction map of the residues involved in the binding of three subunits of the Nt*Eh*CaBP1. A) The interaction amongst the three subunits of N-terminal domain of CaBP1 forming a stable trimer. B) The interactions between the Chain A and Chain B of *Eh*NtCaBP1 are shown in graphical representation. The amino acid involved in the interactions are represented in three letter code. The interacting chains of hexamer and the trimer are joined by colored lines. The different color on different line represents the type of interaction. The area of each circle is proportional to the surface area of the corresponding protein chain. C) The topology one the three α-helix and two calcium binding sites in NtEhCaBP1.

3.4.8 Very small change in overall charge distribution

To understand the differences in the charge distribution in the mutated loop we calculated the overall surface charge distribution of Nt*Eh*CaBP1 EF-II mutant and Nt*Eh*CaBP1(Figure 3.9) using ABPS plugin [25] in PyMol. The analysis suggested that the charge distribution in the calcium binding loop of the mutant remains almost similar to that of Nt*Eh*CaBP1. The minor differences in the calcium binding loops are due to the incorporation of benzyl group of phenylalanine which has a hydrophobic nature due to the presence of the benzyl side chain.

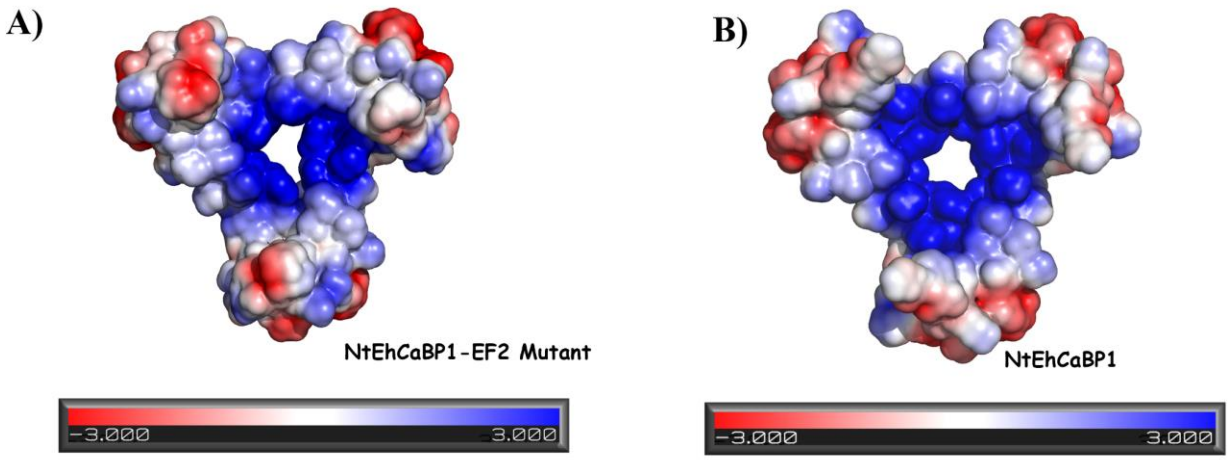
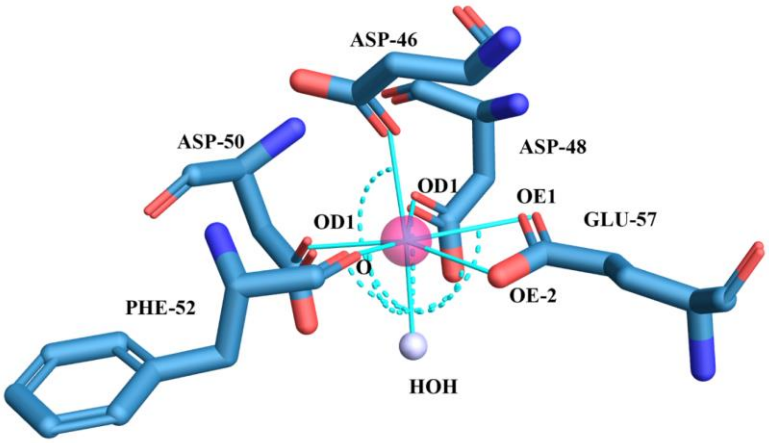


Figure 3.9 The overall charge distribution on the trimer binding interface of Nt*Eh*CaBP1 EF-II mutant and Nt*Eh*CaBP1.

3.4.8 The structural representation and comparison of the NtEhCaBP1 EF-II mutant active site

In order to investigate the mechanism of the designed EF-II mutants high binding calcium affinity, we looked into changes at the atomic level by calculating the distances and angles of the

calcium bound to the residues of the 2nd loop of the mutant protein and comparing the mutant site with the native Nt*Eh*CaBP1.



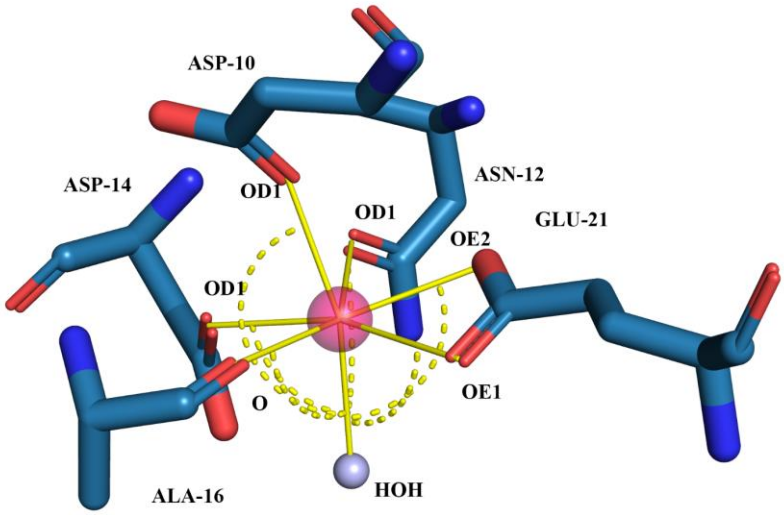
Nt-EhCaBP1-EF2-Mutant

Figure 3.10 The active site coordination of the metal ion at the binding site of the designed mutant is shown.

Residue	AA	Position	Distance	Angle	Location
46A	ASP	X	**2.34**	172.6	HOH-CA-OD1
48A	ASP	Y	2.4	95.7	HOH-CA-OD1
50A	ASP	Z	2.22	84.7	HOH-CA-OD1
52A	PHE	-Y	2.27	97	HOH-CA-O
57A	GLU	-Z	2.66	95.6	HOH-CA-OE1
57A	GLU		2.5	79.5	HOH-CA-OE2
1Z	HOH	-X	**2.32**		water

Table 3.6 The binding residues and water oxygen atoms are labeled in the Figure 3.10 are listed in the table along with atomic distances between the interacting atom/ residue and the changes in the angle position with respect to the water molecule.

The crystal diffracted at 1.9Å resolution showed good electron density in the calcium bound sites. The atomic distances were calculated from the oxygen atoms from the sidechain, main chain and water coordinate the calcium ion in pentagonal bipyramidal geometry.



Nt-EhCaBP1-EF1

Figure 3.11 The active site coordination of the metal ion at the binding site of the designed mutant at 1.9Å resolution is shown in the figure.

Residue	AA	Position	Distance	Angle	Location
46A	ASP	X	2.72	158	HOH-CA-OD1
48A	ASP	Y	2.5	102.4	HOH-CA-OD1
50A	ASN	Z	2.6	73.6	HOH-CA-OD1
52A	GLU	-Y	2.56	105.3	HOH-CA-O
57A	GLU	-Z	2.58	109.1	HOH-CA-OE1

57A	GLU		2.56	115.2	HOH-CA-OE2
75A	HOH	-X	2.92		water

Table 3.7 The binding statistics of the native Nt*Eh*CaBP1-EF2-loop.

The comparison of both sites and the interactions with calcium clearly suggested shrinkage of the coordination sphere (Figure 3.10 and 3.11). The overall shrinkage in the active site was accounted for the difference in the binding affinity. The notable changes were seen in the oxygen atoms from the aspartate residues present in the X, Y, Z positions the water molecule was also closer in the mutant loop. The coordination distance of the main chain oxygen coming from glutamate residue was also notable far compared to the phenyl alanine at the 52nd position with the backbone oxygen atom.

3.4.10 The binding site of EF-1 loop of NtEhCaBP1 EF-II mutant showed tighter binding

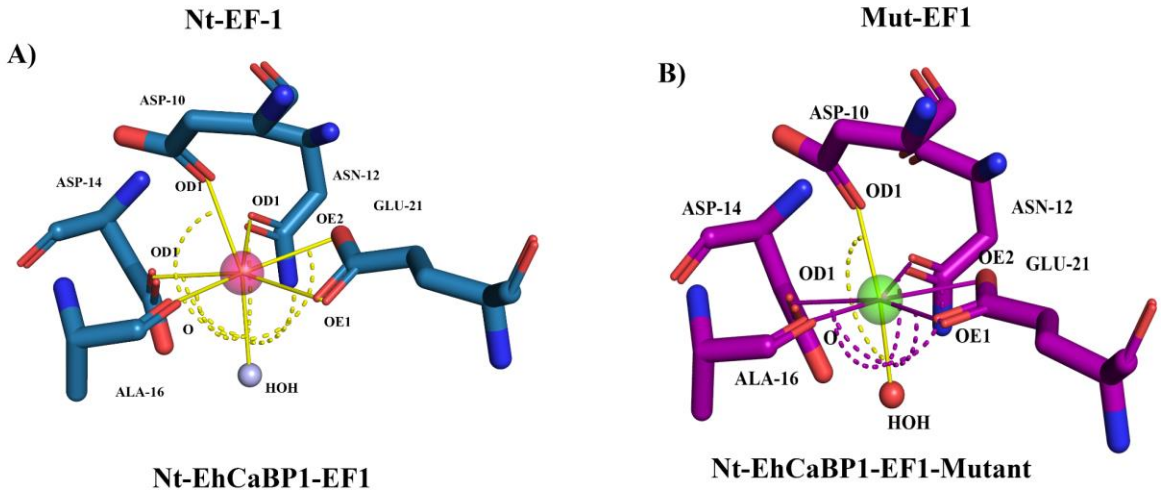


Figure 3.12 The comparison of the binding site and coordinating atoms involved in binding of calcium to the 1st EF-hand motif EF-1 from Nt*Eh*CaBP1 and mutant.

Residue	AA	Position	Distance	Angle	Location
10A	ASP	**X**	**2.11**	166.6	HOH-CA-OD1
12A	ASN	**Y**	2.4	89.9	HOH-CA-OD1
14A	ASP	**Z**	2.58	84.4	HOH-CA-OD1
16A	ALA	**-Y**	**2.25**	96	HOH-CA-O
21A	GLU	**-Z**	2.55	90.6	HOH-CA-OE1
21A	GLU		2.46	93.5	HOH-CA-OE2
208Z	HOH	**-X**	**2.19**		water

Table 3.8 The binding statistics of the Nt*Eh*CaBP1-EF2-mutant loop-I.

Residue	AA	Position	Distance	Angle	Location
10A	ASP	X	**2.5**	**164.1**	HOH-CA-OD1
12A	ASN	Y	2.5	104.1	HOH-CA-OD1
14A	ASP	Z	2.47	83.5	HOH-CA-OD1
16A	ALA	-Y	**2.42**	89	HOH-CA-O
21A	GLU	-Z	2.66	86.3	HOH-CA-OE1
21A	GLU		2.6	107.6	HOH-CA-OE2
71A	HOH	-X	2.66		water

Table 3.9 The binding statistics of the Nt*Eh*CaBP1-loop-I.

The comparative analysis of both the sites, one of which is a low binding site from Nt*Eh*CaBP1 and also the same residues code for the EF-1 loop of the mutant showed shrinkage in the sphere coordinating to calcium (Figure 3.12). We observed one of possible case of cooperative binding that is very much evident from the changes in the atomistic distances of the loop1 of the same protein having same amino acid residues (Table 3.8 and 3.9). The notable changes were seen in the X, Y and –Y positions as well as with the coordinating water molecule.

3.4.11 The curious case of two calcium bound with one EF-hand

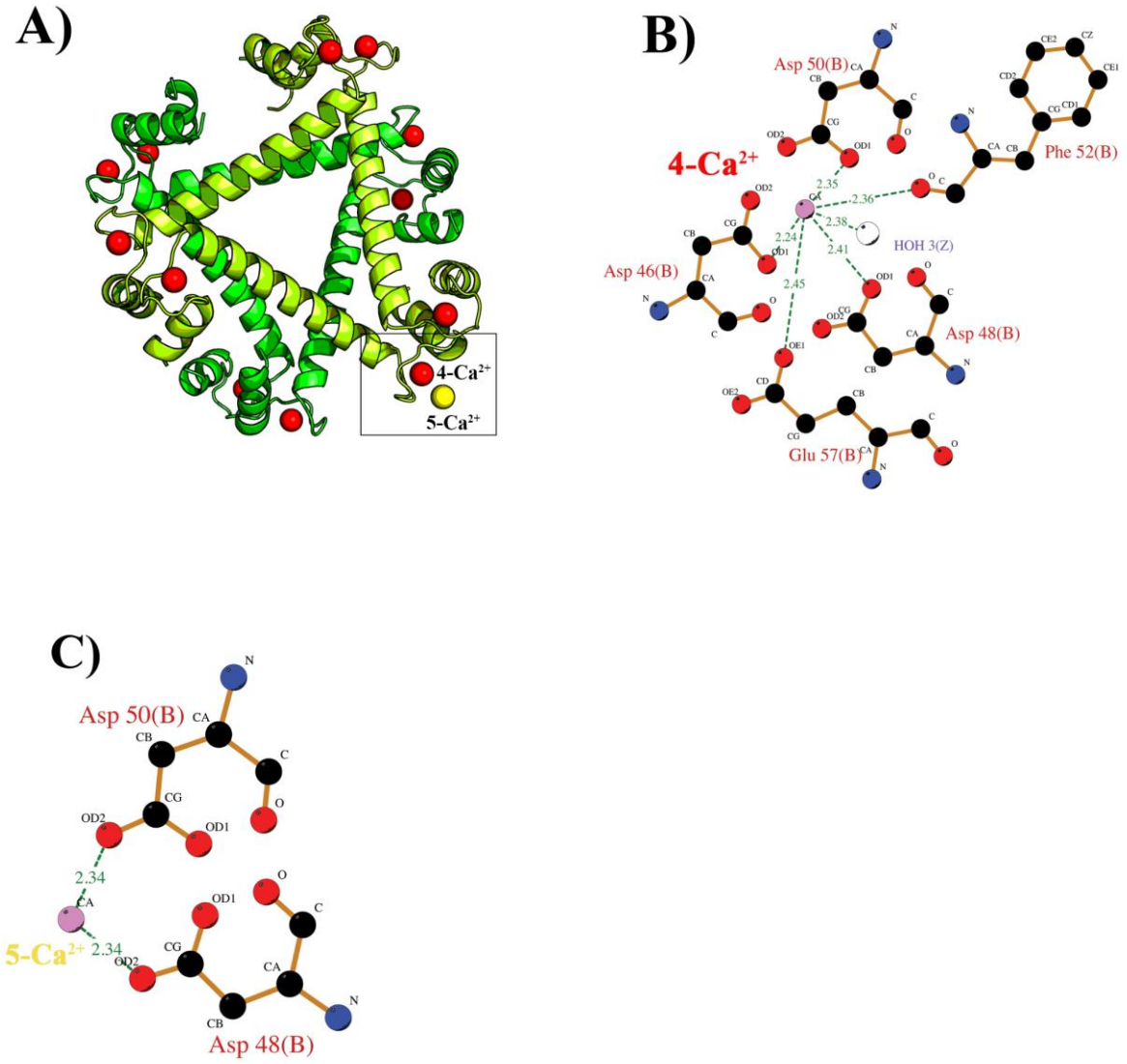


Figure 3.13 Two calcium ions bound to the designed EF-2 of the Nt*Eh*CaBP1-EF2 mutant. **A)** The structure of the mutant protein shown in cartoon representation where the 2nd EF hand of the chain B binds with two calcium ions shown as 4th and 5th calcium ion, the 5th calcium ion positioned away from the site is colored in yellow. B) The ligand plot of all the interactions of the 4th calcium ion with the residues of the mutant EF2 loop is shown. C) The interactions of the 5th calcium ion bound to the residues also involved with the binding of 4th calcium ion are shown.

Interestingly, in the crystal structure of the Nt*Eh*CaBP1-EF2 mutants the second subunit's 2nd EF motif (designed) binds with two calcium ions (Figure 3.13A) compared to all the other loops. In order to understand how the 2nd (5-Ca^{2+}) calcium is binding outside the core EF hand loop where there is already (4- Ca^{2+}) (Figure 3.13B) calcium is bound we calculated and plotted all the atomic interaction involving 4- Ca^{2+} and 5- Ca^{2+} (Figure 3.13C).

The 4- Ca^{2+} interacts with X, Y, Z, -Y, -Z, -X like any other conventional EF-hand loop forming bipyramidal pentagonal geometry and the 5- Ca^{2+} interacts with the 2nd oxygen atom from the aspartate at the Y and Z positions. The aspartates residues at Y and Z position participates in coordinating with calcium via the oxygen at 1st delta position (OD1) which interacts with 4- Ca^{2+} and the oxygen atoms of the OD2 interact with the 5- Ca^{2+}.

3.4.12 The Online and offline services for the use of Cal-EF-Afi2

The updated version of the set of programs (offline) with new scoring functions and a user-friendly webserver (online) to predict, design and engineer EF-hand binding loop is available at http://202.41.10.46/calb/.

3.5 Discussion

The sequence evolution has occurred over the period of millions of years. One of the factors that has been the primary cause of these important events are the mutations in the sequences of many genes. These few to many mutations have been seen in different protein families across all the biological systems. The mutations are mostly caused without altering the structural core of the functional motifs. The EF-hand motif has many conserved residues and the conservation is higher in the calcium binding site. The high preference of certain amino acids such as aspartate at beginning and glutamate at the end clearly indicates that some ligands are indispensable for calcium binding. The presence of the conserved residues in a calcium binding site is important for the calcium binding affinity. These sites have many negatively charged residues. The availability of charged residues in a short stretch of amino acids accounts for an ideal site for the binding of a positively charged divalent ions such as calcium and magnesium. In physiological conditions both

the ions are present in the cell in high concentrations and selectivity of metal ion is highly specific in the EF-hand proteins[32-34].

3.5.1 The unique design of the EF-Loop

In this study, we used SVM margin scores (SVM_{Mar}) and the PSSM based log likelihood scores (PSM_{LogL}) to design a unique calcium binding site which is not present in any of the protein sequence databases. The PSM_{LogL} score is an indicator of the sequence conservation; a high score suggests the predicted site has many residues in the preferred position. Later, we incorporated the designed mutant (DKDGDGFIDFEE) site by using SDMs in second EF-hand loop of Nt*Eh*CaBP1(DADGNGEIDQNE). The design was achieved by performing 5-point mutations incorporated at the following positions: 2nd position: A47K, 5th position: N50D, 7th position: E52F, 10th position: Q55F and 11th position with N56E. The mutations as random as they may appear required replacement of alanine with a positively charged lysine, asparagine at 5th position which is involved in direct interaction with the calcium ion (Z position) is replaced with aspartate followed by the replacement of a negative charged residue glutamic acid at the 7th position (-Y position) and replacement of glutamine with hydrophobic phenylalanine and replacement of asparagine with glutamic acid amongst the non-interacting residues[35-38].

The binding affinities were measured using ITC experiments and compared to the observed native *Eh*CaBP1 binding parameters. The experiments were designed with the aim to modulate the binding affinity of one site and understand the effects by comparing it with the experimentally characterized construct. We observed that the mutations in the 2nd EF-loop of the protein brings ~160 fold and ~523-fold increase in the binding affinity of the calcium binding loops. One of the most important factor that influences the calcium binding affinity is the geometry of the coordination sphere that helps proteins such as EF-hand to bind calcium in a specified geometry (e.g. bipyramidal geometry). In the Nt*Eh*CaBP1-*EF*2-mutant, the enthalpy recorded for the site 1(ΔH, -2415 and 2300 kcal/mol) showed that it is an exothermic reaction with favorable enthalpy and site 2 showed (2300 kcal/mol) unfavorable enthalpy indicating that the binding is an endothermic process. Both the sites in the EF-2 mutant showed favorable entropy of ΔS, ~ 11.0 and 32.8 cal mol^{-1} K^{-1}. It has been shown that calcium ion binding to a site that requires minimal

conformational change is most favorable compared to the site with more flexibility [39]. Tighter binding has shown to be more favorable (entropy). We observed a tighter coordination of ligands in the EF-2 mutant crystal structure that diffracted at 1.9Å resolution. The ability of the protein to provide all Ca^{2+} coordinating oxygen atoms without invoking strain in the polypeptide chain is also an important factor for Ca^{2+} binding affinity. The mutations replacing the small side chains of amino acids such as alanine and asparagine with large side chains of glutamic acid and phenylalanine helps the calcium to bind in a tighter manner.

3.5.3 Oligomerization and high calcium binding affinity

In the Nt*Eh*CaBP1 EF-II mutant, due to the bend in the helix-III, one trimer (interface-A) gets close to the other trimer (interface-B) and forms hexamer. The oligomeric conformation of Nt*Eh*CaBP1 EF-II mutant is formed by a domain swapped trimer very similar to native Nt*Eh*CaBP1, where three symmetry-related molecules interacted in a head-to-tail manner which lead to trimerization of N-terminal domain. The subunits interact amongst each other forming 4 hydrogen bonds compared to 3 in the native trimer structure.

The second subunit of the Nt*Eh*CaBP1 EF-II mutant's second calcium binding site (EF-2) captures two calciums, one interacts with the main chain and side chain oxygen of the five residues of the EF2 loop and the other calcium interacts with the oxygen atoms available with the aspartate from 1st and 3rd positions of the EF-hand loop.

3.5.4 Higher Cooperative Binding in NtEhCaBP1 EF-2 Mutant

The K_a (association constant) calculated after the titrations showed an increase of ~160 fold in calcium binding affinity of site1. The comparison of coordinating distances and angles of EF-1 loop of the Nt*Eh*CaBP1 and Nt*Eh*CaBP1 EF2 mutant showed a shrinkage of the coordination sphere (Table 3.8 and 3.9). The overall shrinkage in the active site was accounted for the difference in the binding affinity.

3.6 Conclusion

We integrated two scoring techniques in the earlier developed method, to design and validate our findings using biochemical, structural and computational techniques. The coordinates obtained after the X-ray diffraction of Nt*Eh*CaBP1 EF2 mutant have been deposited in RCSB protein databank (PDB Code 5XOP). The mutational analysis was carried out by the Cal-EF-Afi2 program. The source of the program is available at http://202.41.10.46/calb/resources.html. The webserver is free accessible for everyone and is available at http://202.41.10.46/calb/. The program is optimal for scanning large protein databases for calcium binding site identification and estimation of binding affinity. The PSM_{LogL} and SVM_{MAR} scores are provided to assist binding affinity modulation for the scientific community working on numerous proteins still to be annotated. The webserver requires only the protein sequence for the prediction without prior knowledge of structural or biochemical information.

3.7 References

1. Gifford Jessica L, Walsh Michael P, Vogel Hans J. Structures and metal-ion-binding properties of the Ca²⁺-binding helix–loop–helix EF-hand motifs. Biochemical Journal. 2007;405(2):199-221. doi: 10.1042/bj20070255.
2. Arruda AP, Hotamisligil GS. Calcium homeostasis and organelle function in the pathogenesis of obesity and diabetes. Cell metabolism. 2015;22(3):381-97. doi: 10.1016/j.cmet.2015.06.010. PubMed PMID: PMC4558313.
3. Liu T, Altman RB. Prediction of calcium-binding sites by combining loop-modeling with machine learning. BMC structural biology. 2009;9:72. doi: 10.1186/1472-6807-9-72. PubMed PMID: 20003365; PubMed Central PMCID: PMC2808310.
4. Strynadka NCJ, James MNG. Crystal Structures of the Helix-Loop-Helix Calcium-Binding Proteins. Annual Review of Biochemistry. 1989;58(1):951-99. doi: 10.1146/annurev.bi.58.070189.004511.
5. Mazumder M, Padhan N, Bhattacharya A, Gourinath S. Prediction and Analysis of Canonical EF Hand Loop and Qualitative Estimation of Ca²⁺ Binding Affinity. PLoS ONE. 2014;9(4):e96202. doi: 10.1371/journal.pone.0096202.
6. Chang C-CaL, Chih-Jen. LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology. 2011;2(3):27:1--:.
7. Bhasin M, Raghava GP. Analysis and prediction of affinity of TAP binding peptides using cascade SVM. Protein Sci. 2004;13(3):596-607. Epub 2004/02/24. doi: 10.1110/ps.03373104
13/3/596 [pii]. PubMed PMID: 14978300; PubMed Central PMCID: PMC2286721.
8. Ramana J, Gupta D. FaaPred: a SVM-based prediction method for fungal adhesins and adhesin-like proteins. PLoS One. 2010;5(3):e9695. Epub 2010/03/20. doi: 10.1371/journal.pone.0009695. PubMed PMID: 20300572; PubMed Central PMCID: PMC2837750.
9. Kumar S, Padhan N, Alam N, Gourinath S. Crystal structure of calcium binding protein-1 from Entamoeba histolytica: a novel arrangement of EF hand motifs. Proteins. 2007;68(4):990-8. Epub 2007/06/08. doi: 10.1002/prot.21455. PubMed PMID: 17554780.
10. Lo SL, Cai CZ, Chen YZ, Chung MC. Effect of training datasets on support vector machine prediction of protein-protein interactions. Proteomics. 2005;5(4):876-84. Epub 2005/02/18. doi: 10.1002/pmic.200401118. PubMed PMID: 15717327.
11. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci U S A. 2000;97(1):262-7. Epub 2000/01/05. PubMed PMID: 10618406; PubMed Central PMCID: PMC26651.
12. Eddy SR. Where did the BLOSUM62 alignment score matrix come from? Nat Biotechnol. 2004;22(8):1035-6. Epub 2004/08/03. doi: 10.1038/nbt0804-1035
nbt0804-1035 [pii]. PubMed PMID: 15286655.
13. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403-10. Epub 1990/10/05. doi: 10.1016/s0022-2836(05)80360-2. PubMed PMID: 2231712.
14. Bhattacharya A, Padhan N, Jain R, Bhattacharya S. Calcium-binding proteins of Entamoeba histolytica. Arch Med Res. 2006;37(2):221-5. Epub 2005/12/29. doi: S0188-4409(05)00335-8 [pii]

10.1016/j.arcmed.2005.10.002. PubMed PMID: 16380322.
15. Otwinowski Z, Minor W. Processing of X-ray diffraction data collected in oscillation mode. Methods in enzymology. 1997;276:307-26. Epub 1997/01/01. PubMed PMID: 27754618.
16. Storoni LC, McCoy AJ, Read RJ. Likelihood-enhanced fast rotation functions. Acta crystallographica Section D, Biological crystallography. 2004;60(Pt 3):432-8. Epub 2004/03/03. doi: 10.1107/s0907444903028956. PubMed PMID: 14993666.
17. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. Acta crystallographica Section D, Biological crystallography. 2004;60(Pt 12 Pt 1):2126-32. Epub 2004/12/02. doi: 10.1107/s0907444904019158. PubMed PMID: 15572765.
18. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. Journal of Applied Crystallography. 1993;26(2):283-91. doi: doi:10.1107/S0021889892009944.
19. Sievers F, Higgins DG. Clustal omega. Current protocols in bioinformatics. 2014;48:3.13.1-6. Epub 2014/12/17. doi: 10.1002/0471250953.bi0313s48. PubMed PMID: 25501942.
20. Tippmann HF. Analysis for free: comparing programs for sequence analysis. Briefings in bioinformatics. 2004;5(1):82-7. Epub 2004/05/22. PubMed PMID: 15153308.
21. Holm L, Laakso LM. Dali server update. Nucleic Acids Res. 2016;44(W1):W351-5. Epub 2016/05/01. doi: 10.1093/nar/gkw357. PubMed PMID: 27131377; PubMed Central PMCID: PMCPMC4987910.
22. Laskowski RA. Protein Structure Databases. Methods in molecular biology (Clifton, NJ). 2016;1415:31-53. Epub 2016/04/27. doi: 10.1007/978-1-4939-3572-7_2. PubMed PMID: 27115626.
23. Laskowski RA, Swindells MB. LigPlot+: multiple ligand-protein interaction diagrams for drug discovery. Journal of chemical information and modeling. 2011;51(10):2778-86. Epub 2011/09/17. doi: 10.1021/ci200227u. PubMed PMID: 21919503.
24. Henikoff S, Greene EA, Pietrokovski S, Bork P, Attwood TK, Hood L. Gene families: the taxonomy of protein paralogs and chimeras. Science. 1997;278(5338):609-14. Epub 1997/10/24. PubMed PMID: 9381171.
25. Holec PV, Hackel BJ. PyMOL360: Multi-user gamepad control of molecular visualization software. Journal of computational chemistry. 2016;37(30):2667-9. Epub 2016/09/21. doi: 10.1002/jcc.24489. PubMed PMID: 27645768.
26. Henikoff JG, Henikoff S. Using substitution probabilities to improve position-specific scoring matrices. Comput Appl Biosci. 1996;12(2):135-43. Epub 1996/04/01. PubMed PMID: 8744776.
27. Xiao X, Shao S, Ding Y, Huang Z, Chou KC. Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. Amino Acids. 2006;30(1):49-54. doi: 10.1007/s00726-005-0225-6.
28. Kumar S, Aslam S, Mazumder M, Dahiya P, Murmu A, Manjasetty BA, et al. Crystal structure of calcium binding protein-5 from Entamoeba histolytica and its involvement in initiation of phagocytosis of human erythrocytes. PLoS pathogens. 2014;10(12):e1004532. Epub 2014/12/17. doi: 10.1371/journal.ppat.1004532. PubMed PMID: 25502654; PubMed Central PMCID: PMCPMC4263763.
29. Aslam S, Bhattacharya S, Bhattacharya A. The Calmodulin-like calcium binding protein EhCaBP3 of Entamoeba histolytica regulates phagocytosis and is involved in actin dynamics.

PLoS pathogens. 2012;8(12):e1003055. Epub 2013/01/10. doi: 10.1371/journal.ppat.1003055. PubMed PMID: 23300437; PubMed Central PMCID: PMCPMC3531509.
30. Gopal B, Swaminathan CP, Bhattacharya S, Bhattacharya A, Murthy MR, Surolia A. Thermodynamics of metal ion binding and denaturation of a calcium binding protein from Entamoeba histolytica. Biochemistry. 1997;36(36):10910-6. Epub 1997/09/09. doi: 10.1021/bi9702546
bi9702546 [pii]. PubMed PMID: 9283081.
31. Garg VK, Avashthi H, Tiwari A, Jain PA, Ramkete PW, Kayastha AM, et al. MFPPI - Multi FASTA ProtParam Interface. Bioinformation. 2016;12(2):74-7. Epub 2017/01/21. doi: 10.6026/97320630012074. PubMed PMID: 28104964; PubMed Central PMCID: PMCPMC5237651.
32. Marsden BJ, Shaw GS, Sykes BD. Calcium binding proteins. Elucidating the contributions to calcium affinity from an analysis of species variants and peptide fragments. Biochemistry and Cell Biology. 1990;68(3):587-601. doi: 10.1139/o90-084.
33. Deng H, Chen G, Yang W, Yang JJ. Predicting calcium-binding sites in proteins - a graph theory and geometry approach. Proteins. 2006;64(1):34-42. Epub 2006/04/18. doi: 10.1002/prot.20973. PubMed PMID: 16617426.
34. Kirberger M, Wang X, Deng H, Yang W, Chen G, Yang JJ. Statistical analysis of structural characteristics of protein Ca2+-binding sites. JBIC Journal of Biological Inorganic Chemistry. 2008;13(7):1169-81. doi: 10.1007/s00775-008-0402-7.
35. Jones LM. Using Protein Design to Understand the Role of Electrostatic Interactions on Calcium Binding Affinity and
Molecular Recognition. Chemistry Dissertations. 2008;Paper 16.
36. Tatusov RL, Altschul SF, Koonin EV. Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. Proc Natl Acad Sci U S A. 1994;91(25):12091-5. Epub 1994/12/06. PubMed PMID: 7991589; PubMed Central PMCID: PMC45382.
37. Wang X, Kirberger M, Qiu F, Chen G, Yang JJ. Towards predicting Ca2+-binding sites with different coordination numbers in proteins with atomic resolution. Proteins. 2009;75(4):787-98. Epub 2008/11/13. doi: 10.1002/prot.22285. PubMed PMID: 19003991; PubMed Central PMCID: PMC2858581.
38. Zeng J, Gao X, Dai Z, Tang B, Tang XF. Effects of metal ions on stability and activity of hyperthermophilic pyrolysin and further stabilization of this enzyme by modification of a Ca2+-binding site. Applied and environmental microbiology. 2014;80(9):2763-72. Epub 2014/02/25. doi: 10.1128/aem.00006-14. PubMed PMID: 24561589; PubMed Central PMCID: PMCPMC3993279.
39. Linse S, Forsen S. Determinants that govern high-affinity calcium binding. Adv Second Messenger Phosphoprotein Res. 1995;30:89-151. Epub 1995/01/01. PubMed PMID: 7695999.

List of figures.

List of Tables.

Appendix III.

Supplementary Table S1. The χ^2 -(chi-square) value for each amino acid residue is estimated with one degree of freedom and significance level P = 0.001. The $\Sigma\chi^2$ values are estimated with 19 degrees of freedom and significance level P < 0.001. The expected (Exp) and observed (Obs) values and the corresponding χ^2 values for amino acid residues and the $\Sigma\chi^2$ values for those positions that do not reach 10.8 and 43.8 (for one and 19 degrees of freedom, respectively) are given more significant.

POSITION	AA RESIDUE	Number of Occ/Obs	Expected	chi-sq
1	E	279	76.73	533.22
	H	25	11.34	16.47
	Q	49	27.62	16.56
	R	50	37.67	4.03
2	A	70	50.08	7.92
	C	25	7.40	41.90
	F	83	65.25	4.83
	I	166	58.24	199.40
	L	302	73.41	711.78
	V	132	43.51	179.93
	W	21	4.29	65.18
3	D	118	102.31	2.41
	E	101	82.88	3.96
	H	17	11.61	2.50

	K	197	60.59	307.12
	Q	71	26.86	72.57
	R	118	35.32	193.51
4	A	72	50.01	9.67
	E	231	78.39	297.11
	K	130	62.90	71.57
	N	50	37.78	3.95
	Q	62	27.17	44.67
	R	75	36.81	39.62
5	A	210	45.24	599.99
	I	176	57.89	240.95
	L	107	80.15	8.99
	M	148	31.38	433.34
	V	100	44.62	68.73
6	F	525	49.98	4514.94
	I	146	58.93	128.65
	M	78	33.80	57.79
	W	31	3.94	185.84
7	A	85	49.56	25.34
	K	158	61.94	148.99
	N	65	37.26	20.66
	Q	73	26.79	79.73
	R	141	34.53	328.32

	S	76	41.75	28.09
8	E	175	80.32	111.59
	H	35	10.99	52.45
	K	79	64.67	3.18
	L	123	79.60	23.67
	M	62	34.36	22.24
	V	74	45.52	17.82
9	A	102	48.98	57.41
	F	247	59.59	589.46
	I	104	60.38	31.51
	L	120	79.70	20.38
	M	54	34.63	10.83
	V	113	44.17	107.25
	Y	92	16.18	355.44
10	D	896	75.42	8928.60
11	A	70	50.08	7.92
	K	208	60.21	362.78
	P	23	3.97	91.07
	Q	50	27.58	18.22
	R	76	36.77	41.84
	T	123	31.69	263.04
	V	77	45.42	21.97
12	D	536	87.86	2285.84

	N	262	30.45	1760.79
13	G	358	45.31	2157.81
	H	37	10.92	62.27
	K	158	61.94	148.99
	N	126	35.15	234.81
	Q	41	27.89	6.16
	R	67	37.09	24.13
14	D	463	90.38	1536.20
	N	146	34.46	361.05
	S	194	37.67	648.68
15	G	805	29.86	20120.37
16	F	105	64.49	25.44
	K	160	61.87	155.66
	Q	67	26.99	59.29
	R	65	37.15	20.87
	T	91	32.80	103.27
	Y	104	15.76	494.03
17	I	559	44.66	5924.32
	L	163	78.22	91.91
	V	152	42.82	278.34
18	D	240	98.09	205.31
	N	96	36.19	98.86
	S	231	36.39	1040.58

	T	163	30.31	580.84
19	F	196	61.35	295.54
	P	35	3.56	277.66
	R	50	37.67	4.03
	V	65	45.83	8.02
	Y	105	15.73	506.79
20	D	191	99.78	83.39
	E	241	78.04	340.26
	N	71	37.05	31.11
	P	17	4.18	39.29
	Q	68	26.96	62.48
	S	60	42.30	7.40
21	D	84	103.48	3.67
	E	755	60.28	8006.97
22	F	415	53.78	2426.20
	L	315	72.96	802.92
	W	16	4.46	29.88
	Y	29	18.35	6.18
23	C	41	6.84	170.48
	K	140	62.56	95.87
	L	108	80.12	9.70
	Q	57	27.34	32.18
	R	107	35.70	142.37

	V	105	44.45	82.49
24	A	112	48.63	82.58
	E	107	82.67	7.16
	H	39	10.85	73.00
	Q	58	27.30	34.51
	R	66	37.12	22.47
	S	65	42.13	12.41
	T	68	33.59	35.23
25	A	162	46.90	282.46
	G	71	55.23	4.50
	I	106	60.31	34.61
	L	147	78.77	59.10
	M	125	32.18	267.76
	V	100	44.62	68.73
26	C	29	7.26	65.13
	I	75	61.38	3.02
	L	274	74.38	535.75
	M	214	29.10	1174.75
	V	89	45.00	43.02
27	A	84	49.60	23.86
	H	24	11.37	14.03
	K	170	61.52	191.27
	Q	55	27.41	27.78

	R	110	35.60	155.49
	S	88	41.34	52.68
	T	72	33.46	44.40
28	A	93	49.29	38.77
	H	27	11.27	21.97
	K	128	62.97	67.15
	N	56	37.57	9.04
	R	74	36.84	37.47
	S	112	40.51	126.18
	T	56	34.01	14.22
29	I	77	61.31	4.01
	L	215	76.42	251.31
	M	56	34.56	13.30
	T	59	33.91	18.57
	V	64	45.86	7.17

Supplementary Table S2. Test Dataset: Summary of EF-hand loops obtained from literature and their macroscopic binding constant along with CAL-EF-AFi predictions (D5). The classification details with supportive binding constants are listed under "Author's Note". (RED colored affinities are the false negatives affinity predictions; Turquoise colored sequences are the false negatives EF-loop predictions)

Protein	EF-Loop Prediction	K_a (M^{-1}) from the whole protein	Authors Notes	Predicted Affinity	Ref.
Parvalbumin *(Cyprinus carpio)*	DQDKSGFIEEDE	K_1= 2.7x10^9 K_2=2.7x10^9	The two metal sites of parvalbumin for Ca^{2+} with equilibrium constants of K_{Ca} = 2.7 X 10 $^9M^{-1}$	High Affinity	
	DSDGDGKIGVDE			High Affinity	[1]
Calmodulin *(Bos taurus)*	DKDGDGTITTKE	K_1 = 1 x 10^7;	Calmodulin contains four relatively high affinity Ca^{2+}sites	High Affinity	
	DADGNGTIDFPE	K_2 = 3.98 x 10^7;		High Affinity	[2]
	DKDGNGYISAAE	K_3 = 3.16 x 10^6;		High Affinity	
	DIDGDGQVNYEE	K_4 = 2.5 x 10^6		High Affinity	
Caltractin *(Chalmydomonas reinhardtii)*	DTDGSGTIDAKE	K_1 = 8.30 x 10^5;	Ca^{2+} binding measurements demonstrated the binding of four Ca^{2+} ions to caltractin with two	High Affinity	
	DKDGSGTIDFEE	K_2 = 8.30 x 10^5,		High Affinity	[3] , [4]

	DDDNSGTITIKD	K_3 = 6.25 x 10^3;	higher affinity and two lower affinity sites.	Low Affinity	
	DRNDDNEIDEDE	K_4 = 6.25 x 10^3		Low Affinity	
calmodulin-like protein *(Homo sapiens)*	DKDGDGCITTRE	K_1=3.80 x 10^5,	Four Ca^{2+}-binding sites. Binding of the first two Ca^{2+}occurs with somewhat higher affinity than that of the last two Ca^{2+}.	High Affinity	[5]
	DRDGNGTVDFPE	K_2=1.90 x 10^5,		Low Affinity	
	DKDGNGFVSAAE	K_3=4.90 x 10^4,		Low Affinity	
	DTDGDGQVNYEE	K_4=1.20 x 10^4		Low Affinity	
Calbindin D9k *(Bos taurus)*	DKNGDGEVSFEE	K_1 = 1.6 x 10^8, K_2 = 4 x 10^8	Ca^{2+} ion binding to calbindin D9k wild type and with different set of mutants.(High Affinity)	Low Affinity	[6]
Calgranulin C (*Sus scrofa)*	DANQDEQVSFKE	K_1=6.50 x 10^4	The protein binds one Ca^{2+}/monomer with a binding constant of about 2 x 10^4,a low affinity site	Low Affinity	[7]
GF14-loop1 *(Arabidopsis)*	ELDTLGEESYKD	K_1=5.50 x 10^4	Low binding affinity	Low Affinity	[8]

			exhibited by GF14 ω.		
Calhepatin *(Lepidosiren paradoxa)*	DKDKSGTLSVDE DTNKDGQVSWQE	K_1=2.90 x 10^5 K_2=6.00 x 10^3	The affinity const-ants determined agree with the fact that S100 protein affinity for Ca^{2+} is low, the affinity of the C-terminal EF-hand being greater than that of the N-terminal EF-hand.	Low Affinity Low Affinity	[9]

Supplementary Table S3. Validation dataset Summary of EF hand loops obtained from ITC studies of CaBPs from *E. histolytica* and their macroscopic binding constant with CAL-EF-AFi's predictions (D7). The classification details with supportive binding constants are listed under "Author's Note" (RED colored affinities are the false positives predictions)

Protein	EF-Loop Prediction	K_a (M^{-1})	Predicted Affinity	Authors Notes	Ref.
ECaBP1 I	DVNGDGAVSYEE	5.25E+03	Low Affinity	EhCaBP1 has one high-affinity site for Ca^{2+} and Mg^{2+}, one high affinity Ca^{2+}-specific site and two low-affinity Ca^{2+}-specific sites	
ECaBP1 II	DADGNGEIDQNE	1.41E+04	High Affinity		
ECaBP1 III	DVDGDGKLTKEE	5.10E+05	High Affinity		
ECaBP1 IV	DANGDGYITLEE	1.55E+06	High Affinity		
ECaBP3 I	DKDNDNKLTAEE	7.28E+04	Low Affinity	One binding site has affinity in the micromolar range, While the other has affinity in the sub-micromolar range.	[10,11] (Table 5)
ECABP3 III	DKEKNGYISASE	4.00E+06	High Affinity		
ECaBP5	DGDGDGYLTLNE	1.18E+07	High Affinity	High Binding site.	
ECaBP6 I	DRDYDGKIDVKQ	1.07E+05	High Affinity	One high affinity Ca^{2+} binding site and one low binding site.	
ECaBP6 II	DQDKDGKIKASD	4.44E+03	Low Affinity		
Ecabp7 I	DKDKSGYLSPDE	9.86E+04	Low Affinity	One high affinity Ca^{2+} binding site and one low binding site.	
Ecabp7 III	DEDGDGKISFQE	1.04E+06	High Affinity		

Supplementary Table S4. Independent dataset (D6) Summary of EF hand loops obtained from Boguta, *et al*., 1988 [12] . The table contains average binding constants of Ca^{2+} for troponin C superfamily (TnC) proteins from experimental data reported by various laboratories. The classification details with supportive binding constants are listed under "Author's Note" (RED colored affinities are the false positives predictions)

Protein	Sequences/Canonical EF loops Predicted	Authors Note	Predicted Affinity	Ref.
Bovine chains αα	I).DEDGDGEVDFQE	Contains low affinity calcium binding sites. The lower affinity calcium-binding sites titrated at a lower pH.	Low Affinity	[13]
Bovine chains αβ	I).DSDGDGECDFQE		Low Affinity	
Human chains β β	I).DNDGDGECDFQE	Six Ca^{2+}-binding sites which assumed to represent three for each β-monomer. Each β subunit was shown to bind one calcium ion with rather high affinity and two other calcium ions with lower affinity.	Low Affinity	[14]
Rat Chain β β	I).DEDGDGECDFQE	Rat brain S100b protein is characterized by two high-affinity Ca^{2+} binding sites with a	Low Affinity	[15]

		K_d of 2 X 10^{-5} M and four lower affinity sites with K_d about 10^{-4} M.		
Frog pI 4-50 **(FPV4- 50)**	II).DQDKSGFIEEDE III).DSDGDGKIGVDE	Muscular parvalbumins from hake proteins have two high affinity sites	High Affinity High Affinity	[16]
Frog pl 4.88	I).DQDQSGFIEKEE II).DKDGDGKIGVDE	Parvalbumins exhibit two independent and equivalent high affinities Ca^{2+}-Mg^{2+} sites.	High Affinity High Affinity	[17]
Pike pl 5.00	I).DADASGFIEEEE	The intrinsic phenyl-alanine and tyrosine fluorescence of pike parvalbumins monitors the binding of Ca^{2+} ions to both their high affinity Ca^{2+} binding CD and EF sites.	High Affinity	[18]
Rabbit (RPV)	I).DKDKSGFIEEEE II).DKDGDGKIGADE	α-parvalbumins from rabbit exhibit two independent and equivalent high-affinity Ca^{2+}-Mg^{2+} sites.	High Affinity High Affinity	[17]
Rat (RTPV)	I).DKDKSGFIEEDE II).DKDGDGKIGVEE	Parvalbumins: Each of their two functional sites binds Ca (II) with an	High Affinity High Affinity	[19]

		affinity of about 10^8 M^{-1}.		
Bovine cardics (BCTNC)	I).LGAEDGCISTKE II).DEDGSGTVDFDE III).DKNADGYIDLEE IV).DKNNDGRIDYDE	The C-terminal peptide contains two Ca^{2+}-binding sites. The third and fourth sites in cardiac-muscle troponin C are represented by the so-called high-affinity Ca^{2+}/Mg^{2+}-binding sites.	Low Affinity Low Affinity High Affinity High Affinity	[20-22]
Amphioxus	I).DYNKDGSIQWED II).DINKDDVVSWEE III).DVSGDGIVDLEE	The two Amphioxus SCP's have three Ca-binding sites of high affinity: two calcium-specific ones and one Ca-Mg site.	**Low Affinity** **Low Affinity** **Low Affinity**	[23]
Nereis	I).DFDKDGAITRMD II).DTNEDNNISRDE III).DTNNDGLLSLEE	Ca2+ the three sites have the same intrinsic affinity (K_a = $1.7x10^8$ M^{-1}) without co-operatively between the sites.	**Low Affinity** **Low Affinity** **Low Affinity**	[24]
Rabbit(RSLC2)	I).DQNRDGIIDKED II).DPEGKGTIKKQF	Myosin contains two DTNB light chains and binds 2 mol of Ca (II) with high affinity.	**Low Affinity** **Low Affinity**	[25-28]
Scallop	I).DVDRDGFVSKDD	Concluded that both RLC-a and RLC-b	Low Affinity	[29]

		bind only one Ca^{2+} with similar affinities to each other.		
Aequorin	I).DVNHNGKISLDE II).DKDQNGAITLDE III).DIDESGQLDVDE	The K_a for one of the two Ca^{2+} is approx. 7x106 M^{-1}	Low Affinity High Affinity Low Affinity	[30,31]
Calcineurin B	I).DLDNSGSLSVEE II).DTDGNGEVDFKE III).DMDKDGYISNGE IV).DKDGDGRISFEE	Demonstrate that calcineurin is also a Ca^{2+}-binding protein with a high affinity for Ca^{2+} (10^{-6} M) in the presence of physio-logical concentrations of Mg^{2+}.	Low Affinity Low Affinity High Affinity High Affinity	[32]
Ca vector protein	I).DANGDGVIDFDE II).DEDGNGVIDIPE	CaVP binds 2 Ca^{2+} atoms in a non-cooperative way with intrinsic binding constant of $8.2x10^6$ forms a high affinity Ca^{2+}-dependent complex.	High Affinity High Affinity	[33]
F. Hepatica FH8	I).DRNGDGKVSAEE II).DKNKDGKLDLKE	FH8 displays low affinity for Ca^{2+}	Low Binder Low Binder	[34]
Human S100A	I).DANHDGRISFDE	Shows weak binding affinity for ca^{2+} One Ca^{2+}-binding site	Low Binder	[35]

Human Polycystin-2	I).DQDGDQELTEHE	with micromolar affinity	Low Binder	[36]
Human Calcineurin	I).DINSDGVLDEQE II).DTNQDRLVTLEE	Ca^{2+} binds with an affinity of 7 µM and causes structural changes. They showed that Ca^{2+} binds to both sites with equal affinity.	Low Binder Low Binder	[37]
Human Centrin3	I).DTDKDEAIDYHE II).DDDDSGKISLRN III).DKDGDGEINQEE	Binds one Ca^{2+} with high and two Ca^{2+} with low affinity.	Low Binder Low Binder High Binder	[38]
Human Centrin2	I).DRDGDGEVSEQE	Binds only one Ca^{2+} per molecule with a significant affinity	High Binder	[39]
S. cerevisiae Centrin	I).DMNNDGFLDYHE II).DDDHTGKISIKN III).DLDGDDEINENE	Cdc31 has one high affinity Ca^{2+} - Mg^{2+} and two lower affinity Ca^{2+} sites.	Low Binder Low Binder High Binder	[40]
Human Calsenilin	I).DINKDGYITKEE II).DRNQDGVVTIEE	Affinities for Ca^{+2} binding at these two sites are greater than 1 µM.	High Binder High Binder	[41]

Supplementary Table S5. Predictions of putative EF-hand containing calcium binding protein and their calcium binding affinities from *E. histolytica* proteome.

Protein	Ca^{2+} Binding Sites	Predicted sites	SVM Scores	Predicted K_a
Cabp8 gi|169802036|gb|eal50453.2| EF-hand calcium-binding domain containing protein	Site I	DEEHTGYIDISE	0.27	Low Affinity
Cabp9 gi|56474642|gb|eal52004.1| EF-hand calcium-binding domain containing protein	Site I	DLDKDGSVNVDE	0.43	Low Affinity
	Site II	DLNDDGEIDIRQ	0.25	Low Affinity
	Site III	DIKDQGKIGAPE	0.14	Low Affinity
	Site IV	DQDLDGFISLKE	0.55	High Affinity
Cabp10 gi|56472778|gb|EAL50237.1| calmodulin, putative	Site I	DADGDKKIECME	0.32	Low Affinity
	Site II	DPEEKGVIDSKE	0.14	Low Affinity
Cabp11 gi|56472561|gb|eal50040.1| calcium-binding protein, putative	Site I	DEDKDGYLKVRE	0.28	High Affinity
	Site II	DQNKIGSITLTQ	0.02	Low Affinity

Cabp12 gi|56474389|gb|eal51761.1| troponin-like protein, putative	Site I	DTDHSGYLDIDE	0.38	Low Affinity
	Site II	DENEDGKMDLNE	0.29	Low Affinity
	Site III	DVNGDGVLDKKE	0.42	Low Affinity
	Site IV	DTDKNGSLDFDE	0.39	Low Affinity
Cabp13 gi|56471188|gb|eal48778.1| calcium-binding protein, putative	Site I	DKDHSGTLEIDE	0.31	Low Affinity
Cabp14 gi|56468461|gb|eal46305.1| EF-hand calcium-binding domain containing protein	Site I	DTDRSGTIEINE	0.49	Low Affinity
	Site II	DVDFNGRISFYE	0.44	High Affinity
	Site III	DTNRSGTMEPHE	0.16	Low Affinity
Cabp15 gi|56470174|gb|eal47854.1| grainin, putative	Site I	DKDKSGTLELNE	0.27	Low Affinity
	Site II	DMDLSGNIGFYE	0.34	Low Affinity
	Site III	DADHSGTMDLNE	0.30	Low Affinity
Cabp16 gi|56466987|gb|eal44984.1| grainin 1	Site I	DKDKSGSLELDE	0.26	Low Affinity
	Site II	DVDLSGSIGFYE	0.35	Low Affinity
	Site III	DKDKSGNLDEQE	0.30	Low Affinity

Cabp18 gi|169802082|gb|eal49043.2| actinin-like protein, putative	Site I	DKDKSGTLELDE	0.30	Low Affinity
	Site II	DADNNGSIGFYE	0.47	Low Affinity
	Site III	DVDQSGSLDITE	0.23	Low Affinity
Cabp19 gi|169800239|gb|eal42646.2| EF-hand calcium-binding domain containing protein	Site I	DRDRSGTLEINE	0.24	Low Affinity
	Site II	DTDFNGHISFYE	0.49	Low Affinity
	Site III	DRNRSGTLEPHE	0.14	Low Affinity
Cabp20 gi|56474807|gb|eal52163.1| calmodulin, putative	Site I	DIDHDKKISRDQ	0.04	Low Affinity
	Site II	DKEENGQIHEAE	0.20	Low Affinity
Cabp21 gi|56467093|gb|eal45078.1| calcineurin b subunit, putative	Site I	DVDNDGFISNPE	0.61	High Affinity
	Site II	DKDRDGKISYEE	0.76	High Affinity
Cabp22 gi|56467590|gb|eal45535.1| EF-hand calcium-binding domain containing protein	Site I	DTNRTGKISFDV	0.18	Low Affinity
	Site II	DVDNDGLLSYEE	0.44	High Affinity
	Site III	DEDNSGSIEGEE	0.45	High Affinity
Cabp23 gi|56465487|gb|eal43738.1| hypothetical protein, conserved	Site I	DINGNGKISKEE	0.63	Low Affinity

	Site II	DLNNDGKIPTDD	0.31	Low Affinity
Cabp24 gi|56472021|gb|eal49541.1| hypothetical protein 40.t00032	Site I	DKDKDELITIEE	0.34	High Affinity
	Site II	DSNNDNKITCKE	0.22	Low Affinity
	Site III	DLNQNGKIDIQE	0.47	Low Affinity
	Site IV	DSDGDNELNLVE	0.11	Low Affinity
	Site V	DIDGSGLIDRME	0.56	High Affinity
Cabp25 gi|56467775|gb|eal45702.1| EF-hand calcium-binding domain containing protein	Site I	DKDKDELITIEE	0.34	High Affinity
	Site II	DSNNDNKITCKE	0.22	Low Affinity
	Site III	DLNQNGKIDIQE	0.47	Low Affinity
	Site IV	DSDGDNELNLVE	0.11	Low Affinity
	Site V	DIDGSGLIDRME	0.56	High Affinity
Cabp26 gi|56474460|gb|eal51827.1| EF-hand calcium-binding domain containing protein	Site I	DSNGDGVLQIDE	0.40	Low Affinity
	Site II	NINCDGYLDKEE	0.10	Low Affinity
	Site III	DGDHDGLINSQE	0.51	High Affinity

Cabp27 gi|56473611|gb|eal51029.1| EF-hand calcium-binding domain containing protein	Site IV	DKDYSKSIEYDE	0.17	Low Affinity
	Site I	DENGDGVLQLDE	0.38	Low Affinity
	Site II	DENGDGVLQLDE	0.38	Low Affinity
	Site III	DDNRNGLIDEDE	0.46	Low Affinity
	Site IV	DTNRDGLLNETE	0.27	Low Affinity

Supplementary Table S7. Calcium binding EF-hand proteins sequences in FASTA format at 60% sequence redundancy with EF-hand loop region RESIDUES LABELLED IN lower case letters. **(D1)** The sequences were taken from Uniprot (Keyword search) / PFAM (alignment) based and then cross validated using RCSB PDB database.

>CABP1_HUMAN/200-227	>GUC1B_BOVIN/96-124
DIEEIIRDVdlngdgrvdfeeFVRMMSR	KLKWTFKIYdkdrngcidrqeLLDIVESI
>CABP1_HUMAN/163-191	>GUC1C_HUMAN/92-120
ELRDAFREFdtngdgeistseLREAMRKL	KLKWYFKLYdadgngsidkneLLDMFMAV
>CABP1_HUMAN/86-114	>KCIP1_HUMAN/137-165
ELREAFREFdkdkdgyincrdLGNCMRTM	KLRWTFNLYdinkdgyinkeeMMDIVKAI
>CALB1_RAT/190-218	>KCIP4_MOUSE/160-188
EFNKAFELYdqdgngyideneLDALLKDL	KLNWAFNLYdinkdgyitkeeMLDIMKAI
>CALB1_RAT/102-130	>MLR_AEQIR/20-48
EFMKTWRKYdtdhsgfieteeLKNFLKDL	EMKEAFSMIdvdrdgfvskedIKAISEQL
>CALBP_ENTHI/1-29	>MLR_PHYPO/6-34
MAEALFKEIdvngdgavsyeeVKAFVSKK	QIQECFQIFdkdndgkvsieeLGSALRSL
>CALL3_HUMAN/121-149	>MLR_TODPA/17-45
EVDEMIRAAdtdgdgqvnyeeFVRVLVSK	ELKEAFTMIdqdrdgfigmedLKDMFSSL
>CALL3_HUMAN/48-76	>NCALD_BOVIN/148-176
ELRDMMSEIdrdgngtvdfpeFLGMMARK	RTEKIFRQMdtnrdgklsleeFIRGAKSD
>CALL3_HUMAN/12-40	>NCALD_BOVIN/64-92
EFKEAFSLF dkdgdgcittreLGTVMRSL	FAEHVFRTFdangdgtidfreFIIALSVT
>CALL3_HUMAN/85-113	>NCALD_BOVIN/100-128

EIREAFRVFdkdgngfvsaaeLRHVMTRL	KLKWAFSMYdldgngyiskaeMLEIVQAI
>CALL5_HUMAN/118-146	>NCS1_HUMAN/64-92
ELDAMIREAdvdqdgrvnyeeFARMLAQE	FATFVFNVFdenkdgriefseFIQALSVT
>CALL5_HUMAN/82-110	>NCS1_HUMAN/100-128
DLQVAFRAFdqdgdghitvdeLRRAMAGL	KLRWAFKLYdldndgyitrneMLDIVDAI
>CALM2_SOYBN/85-113	>NCS1_HUMAN/148-176
ELKEAFRVFdkdqngfisaaeLRHVMTNL	RVDRIFAMMdknadgkltlqeFQEGSKAD
>CALM2_SOYBN/121-149	>NCS1_YEAST/100-128
EVDEMIREAdvdgdgqinyeeFVKVMMAK	KLSWAFELYdlnhdgyitfdeMLTIVASV
>CALM_BOVIN/85-113	>NCS1_YEAST/148-176
EIREAFRVFdkdgngyisaaeLRHVMTNL	RVKKIFKLMdknedgyitldeFREGSKVD
>CALM_BOVIN/121-149	>NCS1_YEAST/64-92
EVDEMIREAdidgdgqvnyeeFVQMMTAK	FANHLFTVFdkdnngfihfeeFITVLSTT
>CALM_PARTE/121-149	>OBL_OBELO/116-142
EVDEMIREAdidgdghinyeeFVRMMVSK	--DAVFDIFdkdgsgtitldeWKAYGKIS
>CALM_PARTE/12-40	>ONCO_RAT/82-109
EFKEAFALFdkdgdgtittkeLGTVMRSL	ETKSLMDAAdndgdgkigadeFQEMVHS-
>CALM_PARTE/85-113	>PDCD6_HUMAN/94-122
ELIEAFKVFdrdgnglisaaeLRHVMTNL	DWQNVFRTYdrdnsgmidkneLKQALSGF
>CALM_YEAST/12-40	>POLC3_CHEAL/47-75
EFKEAFALFdkdnngsissseLATVMRSL	EVRRMMAEIdtdgdgfisfdeFTDFARAN
>CALM_YEAST/85-113	>POLC3_CHEAL/12-40
ELLEAFKVFdkngdglisaaeLKHVLTSI	DRERIFKRFdtngdgkissseLGDALKTL

>CANB1_BOVIN/132-160	>POLC4_BETVE/11-39
IVDKTIINAdkdgdgrisfeeFCAVVGGL	ERERIFKRFdangdgkisaaeLGEALKTL
>CANB1_BOVIN/54-82	>POLC4_BETVE/46-74
LVQRVIDIFdtdgngevdfkeFIEGVSQF	EVKHMMAEIdtdgdgfisfqeFTDFGRAN
>CANB1_BOVIN/91-119	>POLC7_PHLPR/4-32
KLRFAFRIYdmdkdgyisngeLFQVLKMM	DMERIFKRFdtngdgkislseLTDALRTL
>CATR_CHLRE/29-57	>POLC7_PHLPR/39-67
EIREAFDLFdtdgsgtidakeLKVAMRAL	EVQRMMAEIdtdgdgfidfneFISFCNAN
>CATR_CHLRE/102-130	>PRVA_ESOLU/41-69
EILKAFRLFdddnsgtitikdLRRVAKEL	DVKKVFKAIdadasgfieeeeLKFVLKSF
>CATR_CHLRE/138-166	>PRVA_HUMAN/43-71
ELQEMIAEAdrnddneidedeFIRIMKKT	DVKKVFHMLdkdksgfieedeLGFILKGF
>CATR_CHLRE/65-93	>PRVA_TRISE/42-70
EIKKMISEIdkdgsgtidfeeFLTMMTAK	QVKEVFEILdkdqsgfieeeeLKGVLKGF
>CAVP_BRALA/90-118	>PRVB_CYPCA/42-70
EILRAFKVFdangdgvidfdeFKFIMQKV	DVKKAFAIIdqdksgfieedeLKLFLQNF
>CAVP_BRALA/127-155	>Q26068_PLAMG/20-48
EVEEAMKEAdedgngvidipeFMDLIKKS	EMKEAFTMIdqnrdgfidindLKEMFSSL
>CBP_SACER/138-166	>Q39890_SOYBN/121-149
EAAEAFNQVdtngngelsldeLLTAVRDF	EVEQMIKEAdldgdgqvnyeeFVKMMMTV
>CDC31_YEAST/133-161	>Q39890_SOYBN/48-76
ELRAMIEEFdldgdgeineneFIAICTDS	ELQDMISEVdadgngtiefdeFLSLMAKK
>CDC31_YEAST/24-52	>Q39890_SOYBN/12-40

EIYEAFSLFdmnndgfldyheLKVAMKAL	DFKEAFGLFdkdgdgcitveeLATVIRSL
>CDPK1_ARATH/527-555	>Q39890_SOYBN/85-113
HLFAAFTYFdkdgsgyitpdeLQQACEEF	ELKEAFKVFdkdqngyisaseLRHVMINL
>CDPK1_ARATH/455-483	>Q7ZZB9_ONCMY/56-84
GLKEMFNMIdadksgqitfeeLKAGLKRV	ELQEMIDEVdedgsgtvdfdeFLVMMVRC
>CDPK1_ARATH/561-589	>Q868D4_9HEMI/128-156
RIEELMRDVdqdndgridyneFVAMMQKG	DLDAMIDEIdadgsgtvdfeeFMGVMTGG
>CDPK_SOYBN/445-473	>Q868D4_9HEMI/92-120
HIDDMIKEIdqdndgqidygeFAAMMRKG	ELREAFRLYdkegngyistdvMREILAEL
>CDPK_SOYBN/411-439	>Q8WSQ4_PHYPO/56-84
NLVSAFSYFdkdgsgyitldeIQQACKDF	AFNEMFNEAdatgngkiqfpeFLSMMGRR
>CDPK_SOYBN/339-367	>Q9XZV2_EUPOC/28-56
GLKELFKMIdtdnsgtitfdeLKDGLKRV	EIKEAFDLFdtnktgsidyheLKVAMRAL
>CETN2_HUMAN/141-169	>Q9XZV2_EUPOC/64-92
ELQEMIDEAdrdgdgevseqeFLRIMKKT	EILELMNEYdregngyigfddFLDIMTEK
>CETN2_HUMAN/32-60	>RECO_BOVIN/101-129
EIREAFDLFdadgtgtidvkeLKVAMRAL	KLEWAFSLYdvdgngtiskneVLEIVTAI
>CHP1_HUMAN/114-142	>S100B_BOVIN/53-81
KLHFAFRLYdldkdekisrdeLLQVLRMM	VVDKVMETLdsdgdgecdfqeFMAFVAMI
>CHP2_HUMAN/115-143	>S10A1_BOVIN/54-82
KLHYAFQLYdldrdgkisrheMLQVLRLM	AVDKVMKELdengdgevdfqeYVVLVAAL
>CLSS_HAEMA/50-78	>S10AB_PIG/57-85
ASAKLIKMAdknsdgkiskeeFLNANAEL	VLDRMMKKLdldsdgqldfqeFLNLIGGL

>CLSS_HAEMA/8-36	>TNNC1_CHICK/96-124
ELEAAFKKLdangdgyvtaleLQTFMVTL	ELSDLFRMFdknadgyidleeLKIMLQAT
>CSEN_HUMAN/166-194	>TNNC2_CHICK/98-126
KLKWAFNLYdinkdgyitkeeMLAIMKSI	ELANCFRIFdknadgfidieeLGEILRAT
>CSEN_HUMAN/214-242	>TNNC2_CHICK/58-86
HVERFFEKMdrnqdgvvtieeFLEACQKD	ELDAIIEEVdedgsgtidfeeFLVMMVRQ
>CSEN_MOUSE/214-242	>TNNC2_CHICK/134-162
HVERFFQKMdrnqdgvvtideFLETCQKD	DIEDLMKDSdknndgridfdeFLKMMEGV
>GUC1A_CHICK/54-82	>TNNC2_RABIT/95-123
YVEQMFETFdfnkdgyidfmeYVAALSLV	ELAECFRIFdrnadgyidaeeLAEIFRAS
>GUC1B_BOVIN/60-88	>TNNC2_RABIT/131-159
YVEAMFRAFdtngdntidfleYVAALNLV	EIESLMKDGdknndgridfdeFLKMMEGV

Supplementary Table S9. The training data used for estimation of binding affinity were taken from RCSB based on PSSM scores obtained from the EF-hand loop region. The positive dataset (**D3**) consisted of 144, 12-mer sequences and there were 124 sequences in the negative dataset (**D4**).

HIGH BINDERS (D3)	**PSSM SCORE**		**LOW BINDERS (D4)**	**PSSM SCORE**
DRDGDGYISADE	6.77		DADNSGDISLRE	4.89
DKNGDGYIDLEE	6.67		DTDGNGFLDSSE	4.88
DTDGDGYISYQE	6.66		DANNDGRITIDE	4.87
DKDGNGYITVEE	6.55		DSDGNGFLDKSE	4.87
DKDGSGYITVDE	6.56		DQNKSGFIEVEE	4.85
DKDGSGYITLDE	6.54		DADSNGNIEFKE	4.84
DEDGDGYISARE	6.48		DRDRDGEVNVEE	4.84
DKDGSGYITIDE	6.48		DENGDGEVDFQE	4.84
DKDGSGYITIDE	6.48		DFNKDGHIDINE	4.82
DKDGSGYITPDE	6.47		DLNGDGKVDLNE	4.81
DIDGDGYISNGE	6.40		DEDSNGSIDHTE	4.80
DVDGDGYITRSE	6.38		DDNQDGKIDIRE	4.79
DTNGDGYIDRDE	6.38		DADHSGTINSYE	4.79
DKDGDGKIDVDE	6.36		DENKDGAIEFHE	4.78
DVDGDGEIDYEE	6.36		DKDKDGRVNALE	4.77
DLDGNGYISREE	6.31		DKDKNGFLTREE	4.76
DQDGNGYIDENE	6.29		DLDNSGKLDVDE	4.75

DQDGSGYITRDE	6.28		DEDGGGDVDFQE	4.74
DKDGNGYITAQE	6.27		DADHSGKLSFEE	4.72
DVDNDGYITREE	6.25		DKNHDSQIDYEE	4.72
DTDGNGYISFNE	6.23		DNDGSGKLGLKE	4.72
DKDGDGRISFEE	6.22		DKNCDGRLDFDE	4.70
DKDGDGRISFEE	6.22		DLDGNGQVEFPE	4.70
DLDQDGYISQEE	6.22		DRDRDGEVNMDE	4.68
DKDGNGYIEGTE	6.21		DKDKNGELDENE	4.67
DKDGDGKIGVEE	6.18		DEDGQGFIPEDY	4.66
DKNGDGYITVNE	6.19		DKNSDGHVDEDE	4.65
DKDRSGYIEEEE	6.17		DMRNDGAIDFGE	4.64
DKNADGYIDLDE	6.16		DANKDGFVEFDE	4.63
DKNADGYIDGEE	6.15		DASHDGGIDVTE	4.62
DKDGDGKIGVDE	6.13		DEDKSGRLEFEE	4.62
DKDNSGYITKEE	6.13		DENGDGSVNFKE	4.60
DEDGDGKISFEE	6.10		DCDGNGELSNKE	4.59
DKDGNGYILPQE	6.10		DTEGDGVLTVEE	4.59
DKDASGYITIEE	6.07		DENGDGQLSLNE	4.58
DQDGDGRIDYNE	6.06		DADNSGDVDFQE	4.57
DKDGDGKIGIDE	6.05		DGDNDGELEENE	4.56
DKDGDGKISFQE	6.05		DREGQGFISGAE	4.55
DQDKSGYIEEEE	6.05		DKDNSGQVSMKE	4.52
DADKNGYIDFKE	6.04		DKDNDGKVSVED	4.51

DVDGDGVIDYSE	6.05		DKNGTGSVTFDE	4.50
DVDGNGTIDYYE	6.05		DQNRDGFIDKED	4.48
DLNGDGYIQREE	6.04		DEDGDHQVDFKE	4.47
DTDGDGFIDFNE	6.03		DRDHSGTLGPEE	4.47
DKDGDGCITVDE	6.02		DKNSDGTVTWDE	4.46
DKDGDGMIGVDE	6.02		DEKKNGVIEFEE	4.45
DQDGDGFITVEE	6.02		DKNKDRKIDFSE	4.44
DQDKSGYIEEDE	6.01		DKNMDGRLSIDE	4.44
DRDKSGYIEEDE	6.00		DANSDGTLDFKE	4.44
DTDGDGKIGVEE	6.00		DADKDGIIGKND	4.42
DTDGDGKIGVEE	6.00		DINNSGDIDHYE	4.41
DADGDGYVSLQE	5.99		DGDGNSYITTDE	4.41
DMDGDGSIDYLE	5.98		DIDNDGGLNNQE	4.40
DKDGSGAIDFDE	5.98		DTDGTQSIDPKE	4.39
DTDNSGYIEADE	5.98		DTNADGVVDFQE	4.39
DKDGDGKITAAE	5.96		DFDKDGAITRKD	4.39
DKDSSGYITIDE	5.97		DADKSGTMSTYE	4.38
DRNMDGYIDAEE	5.97		DINNDGELTLEE	4.36
DKDASGYISSAE	5.94		DLDKNGKISPDD	4.36
DLDGDGTIDFPE	5.93		DKNADGKLTLQE	4.35
DMDNDGYISNGE	5.93		DSDKSGQLEEKE	4.35
DLDGDGFIDFRE	5.92		DANSDGVVTFDE	4.34
DTDGDGKITSEE	5.93		DANNDGKLSEKE	4.33

DYDRDGTVSLEE	5.93		DKNKDDQITLDE	4.32
DKNEDGYITLDE	5.92		DEDEDGLISRGD	4.31
DADGNGLIDYDE	5.90		DINSDGQLDFQE	4.29
DGDQSGYIEVEE	5.90		DKDNNELIDKQE	4.28
DIDGDGFITPEE	5.91		DKNSDQEIDFKE	4.27
DEDGSGTIDFEE	5.89		DCNNDGQVNYEE	4.26
DEDGSGTIDFEE	5.89		DKDGSRPVDFSE	4.25
DKDGDGKITTKE	5.90		DRDGSRSLDADE	4.25
DTDGDGFISFQE	5.90		DKNGNGTISSLD	4.23
DVDGNGVIDYDE	5.89		DQDGDKQLSLPE	4.22
DADGDGHITFDE	5.88		DIDHNKKIDFTE	4.21
DYDNDGIVSFDE	5.87		DLNSDGEVDMAE	4.20
DLDGSGTIDFEE	5.87		DLNKDNKISWEE	4.19
DTDGDGKISAAE	5.87		DQNRDGFIDIND	4.19
DNDGDGKIGADE	5.85		DINSNGQINLNE	4.16
DADGDGTISFSE	5.85		DEDDSGFITFAN	4.16
DKNGDGFIDKDE	5.85		DPNATGNINKDE	4.15
DRDGDGEINEEE	5.85		DQRGNHQIDFDE	4.14
DVDGDGHISQEE	5.85		DINRSGFVDFTE	4.12
DVDNDGYLDYGE	5.85		DVNCDGRMQFDE	4.12
DKDNDGRIDYSE	5.82		DGNHDGGLNREE	4.11
DADGNGEIDFEE	5.81		DVDRSGTMNSYE	4.11
DNDNSGYITMEE	5.81		DLNKDGVLSRSE	4.10

DFNKDGYIDFME	5.81		DPNRDGHVSLQE	4.07
DKDQNGYISPSE	5.80		DVDRDGFVNKDD	4.07
DRDGDGFISPAE	5.80		DMNNDGRMDQLE	4.05
DSDGDGAITEDE	5.80		DINTDGAVNFQE	4.04
DTDGDGVINYEE	5.81		DVNSDNAINFEE	4.02
DTDKDGKISYEE	5.80		DGNGDGFVCFDD	4.01
DVDGDGQINYEE	5.81		DINSDNAINFEE	4.01
DKDEDGKISFDE	5.79		DTSGSGMIDLND	3.99
DKDGSGTIDTKE	5.78		DKNRTGRLSPEE	3.98
DKDGNGTISKDE	5.77		DLNDDGRVQFNE	3.98
DADGDGMIGIDE	5.76		DCDRDGLVTYDD	3.97
DLNHDGYITFDE	5.75		DKDNDRFVTKCE	3.97
DKNGDGKISVDE	5.74		DSNKNGTLDPSE	3.95
DKDGDGAITRSE	5.73		DFDDDGTLNRED	3.94
DFDGDGMINYEE	5.72		DLNQDGVLTSQE	3.93
DVDKDGYLDVNE	5.72		DADKDGVVTVND	3.92
DEDGSGKIEFEE	5.70		DKDGNNTMNIKE	3.88
DKNKSGYIEIEE	5.70		DTNRSGTITYEQ	3.87
DLDKDGKISFEE	5.70		DANGDNKLDQLE	3.85
DFDKNGYIEYSE	5.70		DKNKDDKLTFDE	3.85
DKNGDGLISVEE	5.68		EQDHDGRVDFFE	3.83
DKDGNGFISAAE	5.68		DQDKSDFVEEDE	3.82
DIDGNGKISVEE	5.66		DQNRDGIICKAD	3.81

DKDGSGHITKEE	5.67		DSNCSGTLSKKE	3.81
DVDGNGSIDYVE	5.65		DLNKNGQVELNE	3.78
DANGDGVIDFDE	5.65		DRDDDGVVSRGD	3.76
DEDGSGEIEFEE	5.63		DRNASDTISCDE	3.75
DEDGSGQIEFEE	5.62		DKNNDDLLSVDE	3.73
DEDGSGTIDFNE	5.62		DRNRSGTLEPHE	3.73
DKDQDGLISKDE	5.60		DKNNDAQLTLEE	3.69
DQDNDGRIDYGE	5.61		DKNKDNKMSFKE	3.68
DSDNDGRIDYSE	5.60		DVNHDGVVSFDD	3.68
DIDGDGQITSKE	5.58		DKNNDEAVDKKE	3.66
DKDCDGNIDFQE	5.58		DIDNNGFLDQND	3.65
DCDGDGKINRKE	5.56		DTNSDGKVEEDD	3.63
DKDGNGTISIKE	5.56		DTNQDNQLSFEE	3.62
DANGDGYFTLEE	5.56		DCNKDNEVDFQE	3.55
DVDGNGKIDFGE	5.56		DANQDEQVDFQE	3.52
DIDGNGTIDEKE	5.55			
DKNGDGRITKEE	5.55			
DADGSGYLEGKE	5.53			
DTDGDGKIAPSE	5.53			
DKDNSGYLTVDE	5.53			
DKDKDGFIEKME	5.52			
DADEKGYIEEKE	5.50			
DLDNDGKIDFSE	5.51			

DKDGDGCVTVEE	5.49	
DIDGSGSIDASE	5.48	
DHDRDGFISQEE	5.47	
DKDGNGLITAAE	5.47	
DLDQDGRISFDE	5.45	
DEDGSGTIDPVE	5.45	
DADGNGSIDKNE	5.44	
DKDKNGKISPEE	5.42	
DHDHDGYISQED	5.42	
DTNGDGSIDFRE	5.41	
DRDNDGYLSDTE	5.39	
DANGDGKISAAE	5.38	

Supplementary Table S10. The redundant set of PDB IDs of EF-hand containing calcium binding proteins. The sequences taken from RCSB were further processed using CD-HIT and the list if the sequences with different threshold are listed in supplementary sheet 5.

1A03; 1A29; 1A2X; 1A75; 1AHR; 1AJ4; 1AJ5; 1AK8; 1ALV; 1ALW; 1AP4; 1AUI; 1AVS; 1B1G; 1B4C; 1B7T; 1B8C; 1B8L; 1B8R; 1B9A; 1BJF; 1BLQ; 1BMO; 1BOC; 1BOD; 1BU3; 1C07; 1C7V; 1C7W; 1CB1; 1CDL; 1CDM; 1CDN; 1CDP; 1CFC; 1CFD; 1CFF; 1CFP; 1CKK; 1CLB; 1CLL; 1CLM; 1CM1; 1CM4; 1CMF; 1CMG; 1CNP; 1CTA; 1CTD; 1CTR; 1DEG; 1DF0; 1DFK; 1DFL; 1DGU; 1DGV; 1DJG; 1DJH; 1DJI; 1DJW; 1DJX; 1DJY; 1DJZ; 1DMO; 1DT7; 1DTL; 1DVI; 1EH2; 1EJ3; 1EL4; 1EXR; 1F4O; 1F4Q; 1F54; 1F55; 1F70; 1F71; 1F8H; 1FF1; 1FI5; 1FI6; 1FPW; 1FW4; 1G33; 1G4Y; 1G8I; 1GGW; 1GGZ; 1GJY; 1H4B; 1HQV; 1HT9; 1I84; 1IG5; 1IGV; 1IH0; 1IJ5; 1IJ6; 1IKU; 1IQ3; 1IQ5;

1IRJ; 1IWQ; 1J1D; 1J1E; 1J7O; 1J7P; 1JBA; 1JC2; 1JF0; 1JF2; 1JFJ; 1JFK; 1JSA; 1JUO; 1JWD; 1K2H; 1K8U; 1K90; 1K93; 1K94; 1K95; 1K96; 1K9K; 1K9P; 1K9U; 1KCY; 1KFU; 1KFX; 1KK7; 1KK8; 1KQM; 1KQV; 1KSM; 1KWO; 1L2O; 1L7Z; 1LA0; 1LA3; 1LIN; 1LKJ; 1LVC; 1LXF; 1M31; 1M39; 1M63; 1M8Q; 1MF8; 1MHO; 1MQ1; 1MR8; 1MUX; 1MVW; 1MWN; 1MXE; 1MXL; 1N0Y; 1N65; 1NCX; 1NCY; 1NCZ; 1NIW; 1NP8; 1NPQ; 1NSH; 1NUB; 1NWD; 1NX0; 1NX1; 1NX2; 1NX3; 1NYA; 1O18; 1O19; 1O1A; 1O1B; 1O1C; 1O1D; 1O1E; 1O1F; 1O1G; 1OHZ; 1OMD; 1OMR; 1OMV; 1OOJ; 1OQP; 1OSA; 1OZS; 1PAL; 1PK0; 1PON; 1PRW; 1PSB; 1PSR; 1PVA; 1PVB; 1Q80; 1QIV; 1QIW; 1QLK; 1QLS; 1QV0; 1QV1; 1QVI; 1QX2; 1QX5; 1QX7; 1QXP; 1REC; 1RFJ; 1RJV; 1RK9; 1RRO; 1RTP; 1RWY; 1S1E; 1S26; 1S36; 1S3P; 1S5G; 1S6C; 1S6I; 1S6J; 1SBJ; 1SCM; 1SCV; 1SK6; 1SKT; 1SL7; 1SL8; 1SL9; 1SMG; 1SNL; 1SPY; 1SR6; 1SRA; 1SW8; 1SY9; 1SYM; 1TCF; 1TCO; 1TIZ; 1TN4; 1TNP; 1TNQ; 1TNW; 1TNX; 1TOP; 1TRF; 1TTX; 1U5I; 1UHH; 1UHI; 1UHJ; 1UHK; 1UP5; 1UWO; 1WDC; 1WRK; 1WRL; 1WRZ; 1X02; 1XA5; 1XFU; 1XFV; 1XFW; 1XFX; 1XFY; 1XFZ; 1XK4; 1XO5; 1XVJ; 1XYD; 1Y0V; 1Y1A; 1Y6W; 1YR5; 1YRT; 1YRU; 1YTZ; 1YV0; 1YX7; 1YX8; 1ZAC; 1ZFS; 1ZMZ; 1ZOT; 1ZUZ; 2A4J; 2AAO; 2AMI; 2B1U; 2B59; 2BBM; 2BBN; 2BCA; 2BCB; 2BCX; 2BE4; 2BE6; 2BEC; 2BKH; 2BKI; 2BL0; 2CCL; 2CNP; 2COL; 2CT9; 2CTN; 2D8N; 2DFS; 2DOQ; 2E6W; 2F2O; 2F2P; 2F33; 2F3Y; 2F3Z; 2F8P; 2FOT; 2G9B; 2GGM; 2GGZ; 2GV5; 2H61; 2HET; 2HF5; 2HPS; 2HQ8; 2HQW; 2I08; 2I18; 2I2R; 2I94; 2ISD; 2IX7; 2JC2; 2JPT; 2JQ6; 2JT0; 2JT3; 2JT8; 2JTT; 2JTZ; 2JU0; 2JUL; 2JWW; 2JXC; 2JXL; 2JZI; 2K0E; 2K0F; 2K0J; 2K2F; 2K2I; 2K3S; 2K61; 2K7B; 2K7C; 2K7D; 2K7O; 2KAX; 2KAY; 2KBM; 2KDH; 2KDU; 2KFF; 2KFG; 2KFH; 2KFX; 2KGB; 2KGR; 2KHN; 2KNE; 2KQY; 2KRD; 2KSP; 2KUG; 2KUH; 2KXW; 2KYC; 2KYF; 2KZ2; 2L0P; 2L1R; 2L2E; 2L4H; 2L4I; 2L50; 2L51; 2L53; 2L7L; 2L98; 2LAN; 2LAP; 2LCP; 2LGF; 2LHH; 2LHI; 2LHL; 2LL6; 2LL7; 2LLO; 2LLQ; 2LLS; 2LLT; 2LLU; 2LM5; 2LMT; 2LMU; 2LMV; 2LNK; 2LP2; 2LP3; 2LQC; 2LQP; 2LUC; 2LUX; 2LV6; 2LV7; 2LVI; 2LVJ; 2LVK; 2LVV; 2M3S; 2M55; 2M7K; 2M7M; 2M7N; 2MA2; 2MAZ; 2MYS; 2NLN; 2NXQ; 2NZ0; 2O5G; 2O60; 2OBH; 2OPO; 2P6B; 2PAL; 2PAS; 2PMY; 2PQ3; 2PRU; 2PSR; 2PVB; 2Q4U; 2Q91; 2QPT; 2R28; 2R2I; 2RGI; 2RO9; 2RRT; 2SAS; 2SCP; 2TN4; 2V01; 2V02; 2V53; 2VAS; 2VAY; 2VB6; 2VN5; 2VN6; 2VRG; 2W49; 2W490; 2W4A; 2W4G; 2W4H; 2W4T; 2W4U; 2W4U0; 2W4V; 2W4W; 2W73; 2WEL; 2WND; 2WOR; 2WOS; 2X0G; 2X51; 2Y4V; 2YGG; 2ZN8; 2ZN9; 2ZND; 2ZNE; 2ZRS; 2ZRT; 3A4U; 3A8R; 3AAJ; 3AAK; 3B32; 3BOW; 3BXK; 3BXL; 3BYA; 3C1V; 3CGA; 3CLN; 3CR2; 3CR4; 3CR5; 3CS1; 3CTN; 3CZT; 3D0Y; 3D10; 3DD4; 3DF0; 3DVE; 3DVJ; 3DVK; 3DVM; 3E3R; 3EK4; 3EK7; 3EK8; 3EKH; 3EVR; 3EVU; 3EVV; 3EWT; 3EWV; 3F45; 3FS7; 3FWB; 3FWC; 3G43; 3GK1; 3GK2; 3GK4; 3GN4; 3GOF; 3GP2; 3H4S; 3HCM; 3HR4; 3I5F; 3I5G; 3I5H; 3I5I; 3ICB; 3IF7; 3IFK; 3IQO; 3IQQ; 3J04; 3J41; 3JTD; 3JVT; 3K21; 3KCP; 3KF9; 3KO0; 3L9I; 3LCP; 3LI6; 3LK0; 3LK1; 3LL8; 3LLE; 3M0W; 3NXA; 3O77; 3O78; 3OX5; 3OX6; 3OXQ; 3PAL; 3PAT; 3PM8; 3PSR; 3PX1; 3QJK; 3QRX; 3RLZ; 3RM1; 3RV5;

3SG2; 3SG3; 3SG4; 3SG5; 3SG6; 3SG7; 3SJQ; 3SUI; 3UCT; 3UCW; 3UCY; 3ULG; 3WFN; 3ZWH; 4ANJ; 4AQI; 4AQJ; 4AQR; 4CLN; 4CPV; 4DBP; 4DBQ; 4DCK; 4DIR; 4DJC; 4DS7; 4DUQ; 4E50; 4E53; 4EHQ; 4ETO; 4F0Z; 4FL4; 4FQO; 4G27; 4G28; 4GGF; 4GOW; 4HEX; 4HSZ; 4I2Y; 4I5J; 4I5K; 4I5L; 4I5N; 4ICB; 4IL1; 4J9Y; 4J9Z; 4L9M; 4PAL; 4TNC; 5CPV; 5PAL; 5TNC

Supplementary Table S11.

Protein	EF-Loop Prediction	K_a (M^{-1}) from the whole protein	Predicted Affinity	SVM_{Mar}	PSM_{LogL}
Parvalbumin *Cyprinus carpio*	DQDKSGFIEEDE DSDGDGKIGVDE	K1= $2.7x10^9$ K2=$2.7x10^9$	High Affinity High Affinity	0.378 1.956	5.252 5.884
Calmodulin *Bos taurus*	DKDGDGTITTKE DADGNGTIDFPE DKDGNGYISAAE DIDGDGQVNYEE	K1 = 1×10^7; K2 = 3.98×10^7; K3 = 3.16×10^6; K4 = 2.5×10^6	High Affinity High Affinity High Affinity High Affinity	1.886 0.937 1.871 0.145	5.874 5.589 6.398 5.19
Caltractin *Chalmydomonas reinhardtii*	DTDGSGTIDAKE DKDGSGTIDFEE DDDNSGTITIKD DRNDDNEIDEDE	K1 = 8.30×10^5; K2 = 8.30×10^5, K3 = 6.25×10^3; K4 = 6.25×10^3	High Affinity High Affinity Low Affinity Low Affinity	1.119 1.205 -1.483 -1.351	5.626 6.094 4.059 4.215
Calmodulin-like protein *Homo sapiens*	DKDGDGCITTRE DRDGNGTVDFPE DKDGNGFVSAAE DTDGDGQVNYEE	K1=3.80×10^5, K2=1.90×10^5, K3=4.90×10^4, K4=1.20×10^4	High Affinity Low Affinity Low Affinity Low Affinity	1.064 -0.587 -0.266 -0.051	5.675 4.995 5.089 5.226
Calbindin D9k *Bos taurus*	DKNGDGEVSFEE Non-canonical Site	K1 = 1.6×10^8, K2 = 4×10^8	Low Affinity **NA**	-1.351 **NA**	5.17 **NA**

Calgranulin C *Sus scrofa*	DANQDEQVSFKE	K1=6.50 x 10^4	Low Affinity	-3.537	3.339
Calhepatin *Lepidosiren paradoxa*	DKDKSGTLSVDE DTNKDGQVSWQE	K1=2.90 x 10^5 K2=6.00 x 10^3	Low Affinity Low Affinity	-1.201 -2.366	4.87 4.199

Protein	**Sequences/Canonical EF loops**	**Experimental Classification**	**Predicted Affinity**	**SVM_{Mar}**	**PSM_{LogL}**
Bovine chains αα	DEDGDGEVDFQE		Low Affinity	-0.35	5.301
Bovine chains αβ	DSDGDGECDFQE		Low Affinity	-0.836	5.192
Human chains β β	DNDGDGECDFQE		Low Affinity	-0.760	5.15
Rat Chain β β	DEDGDGECDFQE		Low Affinity	-0.785	5.192
Frog pI 4-50 (FPV4- 50)	DQDKSGFIEEDE DSDGDGKIGVDE		High Affinity High Affinity	0.378 1.956	5.252 5.884
Frog pl 4.88	DQDQSGFIEKEE DKDGDGKIGVDE		High Affinity High Affinity	0.555 2.267	5.169 6.108
Pike pl 5.00	DADASGFIEEEE		High Affinity	0.890	5.223
Rabbit (RPV)	DKDKSGFIEEEE DKDGDGKIGADE DKDKSGFIEEDE DKDGDGKIGVEE		High Affinity High Affinity High Affinity High Affinity	0.499 2.378 0.484 2.206	5.496 6.096 5.446 6.158
Rat (RTPV)					

Bovine cardics (BCTNC)	LGAEDGCISTKE DEDGSGTVDFDE DKNADGYIDLEE DKNNDGRIDYDE		Low Affinity Low Affinity High Affinity High Affinity	-3.552 -0.578 1.268 0.571	3.398 5.244 6.185 5.551
Amphioxus	DYNKDGSIQWED DINKDDVVSWEE DVSGDGIVDLEE		Low Affinity Low Affinity Low Affinity	-2.164 -2.437 -0.506	4.976 3.535 4.770
Nereis	DFDKDGAITRMD DTNEDNNISRDE DTNNDGLLSLEE		Low Affinity Low Affinity Low Affinity	-0.849 -1.880 - 0.705	4.299 4.0001 4.435
Rabbit(RSLC2)	DQNRDGIIDKED DPEGKGTIKKQF		Low Affinity Low Affinity	-1.831 -3.750	4.33 3.341
Scallop	DVDRDGFVSKDD		Low Affinity	-1.445	4.175
Aequorin	DVNHNGKISLDE DKDQNGAITLDE DIDESGQLDVDE		Low Affinity High Affinity Low Affinity	-0.882 0.365 -0.896	4.641 5.143 4.613

Calcineurin B	DLDNSGSLSVEE DTDGNGEVDFKE DMDKDGYISNGE DKDGDGRISFEE		Low Affinity Low Affinity High Affinity High Affinity	-1.311 -0.992 1.972 2.325	4.562 4.988 5.965 6.196
Ca vector protein	DANGDGVIDFDE DEDGNGVIDIPE		High Affinity High Affinity	0.707 0.806	5.627 5.329
F. Hepatica FH8	DRNGDGKVSAEE DKNKDGKLDLKE		Low Binder Low Binder	-0.680 -1.711	4.98 4.677
Human S100A	DANHDGRISFDE		Low Binder	0.178	5.017
Human Polycystin-2	DQDGDQELTEHE		Low Binder	-1.428	4.068
Human Calnuc	DINSDGVLDEQE DTNQDRLVTLEE		Low Binder Low Binder	-0.998 -2.638	4.19 3.462
Human Centrin3	DTDKDEAIDYHE DDDDSGKISLRN DKDGDGEINQEE		Low Binder Low Binder High Binder	-1.092 -0.240 1.752	4.754 3.984 5.948
Human Centrin2	DRDGDGEVSEQE		High Binder	0.121	5.121

S. cerevisiae Centrin	DMNNDGFLDYHE DDDHTGKISIKN DLDGDDEINENE		Low Binder Low Binder High Binder	-0.346 -0.710 0.145	4.539 3.802 4.638
Human Calsenilin	DINKDGYITKEE DRNQDGVVTIEE		High Binder High Binder	0.308 -1.121	5.801 4.268

Supplementary References:

1. Moeschler HJ, Schaer JJ, Cox JA (1980) A thermodynamic analysis of the binding of calcium and magnesium ions to parvalbumin. Eur J Biochem 111: 73-78.

2. Linse S, Helmersson A, Forsen S (1991) Calcium binding to calmodulin and its globular domains. J Biol Chem 266: 8050-8054.

3. Weber C, Lee VD, Chazin WJ, Huang B (1994) High level expression in Escherichia coli and characterization of the EF-hand calcium-binding protein caltractin. J Biol Chem 269: 15795-15802.

4. Veeraraghavan S, Fagan PA, Hu H, Lee V, Harper JF, et al. (2002) Structural independence of the two EF-hand domains of caltractin. J Biol Chem 277: 28564-28571.

5. Rhyner JA, Koller M, Durussel-Gerber I, Cox JA, Strehler EE (1992) Characterization of the human calmodulin-like protein expressed in Escherichia coli. Biochemistry 31: 12826-12832.

6. Linse S, Johansson C, Brodin P, Grundstrom T, Drakenberg T, et al. (1991) Electrostatic contributions to the binding of Ca2+ in calbindin D9k. Biochemistry 30: 154-162.

7. Dell'Angelica EC, Schleicher CH, Santome JA (1994) Primary structure and binding properties of calgranulin C, a novel S100-like calcium-binding protein from pig granulocytes. J Biol Chem 269: 28929-28936.

8. Lu G, Sehnke PC, Ferl RJ (1994) Phosphorylation and calcium binding properties of an Arabidopsis GF14 brain protein homolog. Plant Cell 6: 501-510.

9. Di Pietro SM, Santome JA (2002) Structural and biochemical characterization of calhepatin, an S100-like calcium-binding protein from the liver of lungfish (Lepidosiren paradoxa). Eur J Biochem 269: 3433-3441.

10. Gopal B, Swaminathan CP, Bhattacharya S, Bhattacharya A, Murthy MR, et al. (1997) Thermodynamics of metal ion binding and denaturation of a calcium binding protein from Entamoeba histolytica. Biochemistry 36: 10910-10916.

11. Rout AK, Padhan N, Barnwal RP, Bhattacharya A, Chary KV (2010) Calmodulin-like Protein from Entamoeba histolytica: Solution Structure and Calcium-Binding Properties of a Partially Folded Protein. Biochemistry.

12. Boguta G, Stepkowski D, Bierzynski A (1988) Theoretical estimation of the calcium-binding constants for proteins from the troponin C superfamily based on a secondary structure prediction method. I. Estimation procedure. J Theor Biol 135: 41-61.

13. Baudier J, Glasser N, Gerard D (1986) Ions binding to S100 proteins. I. Calcium- and zinc-binding properties of bovine brain S100 alpha alpha, S100a (alpha beta), and S100b (beta beta) protein: Zn2+ regulates Ca2+ binding on S100b protein. J Biol Chem 261: 8192-8203.

14. Baudier J, Glasser N, Haglid K, Gerard D (1984) Purification, characterization and ion binding properties of human brain S100b protein. Biochim Biophys Acta 790: 164-173.

15. Baudier J, Labourdette G, Gerard D (1985) Rat brain S100b protein: purification, characterization, and ion binding properties. A comparison with bovine S100b protein. J Neurochem 44: 76-84.

16. Benzonana G, Capony JP, Pechere JF (1972) The binding of calcium to muscular parvalbumins. Biochim Biophys Acta 278: 110-116.

17. Haiech J, Derancourt J, Pechere JF, Demaille JG (1979) Magnesium and calcium binding to parvalbumins: evidence for differences between parvalbumins and an explanation of their relaxing function. Biochemistry 18: 2752-2758.

18. Permyakov EA, Medvedkin VN, Kalinichenko LP, Burstein EA (1983) Comparative study of physiochemical properties of two pike parvalbumins by means of their intrinsic tyrosyl and phenylalanyl fluorescence. Arch Biochem Biophys 227: 9-20.

19. Williams TC, Corson DC, Oikawa K, McCubbin WD, Kay CM, et al. (1986) 1H NMR spectroscopic studies of calcium-binding proteins. 3. Solution conformations of rat apo-alpha-parvalbumin and metal-bound rat alpha-parvalbumin. Biochemistry 25: 1835-1846.

20. Burtnick LD, Kay CM (1977) The calcium-binding properties of bovine cardiac troponin C. FEBS Lett 75: 105-110.

21. Leavis PC, Kraft EL (1978) Calcium binding to cardiac troponin C. Arch Biochem Biophys 186: 411-415.

22. Barskaya NV, Gusev NB (1982) Biological activities of bovine cardiac-muscle troponin C C-terminal peptide (residues 84-161). Biochem J 207: 185-192.

23. Takagi T, Konishi K, Cox JA (1986) Amino acid sequence of two sarcoplasmic calcium-binding proteins from the protochordate Amphioxus. Biochemistry 25: 3585-3592.

24. Cox JA, Stein EA (1981) Characterization of a new sarcoplasmic calcium-binding protein with magnesium-induced cooperativity in the binding of calcium. Biochemistry 20: 5430-5436.

25. Bagshaw CR (1977) On the location of the divalent metal binding sites and the light chain subunits of vertebrate myosin. Biochemistry 16: 59-67.

26. Sugden EA, Nihei T (1969) The effects of calcium and magnesium ions on the adenosine triphosphatase and inosine triphosphatase activities of myosin A. Biochem J 113: 821-827.

27. Okamoto Y, Yagi K (1977) Inhibition by Mg2+ of the interaction of Ca2+ with spin-labeled g2 bound to myosin. J Biochem 82: 835-837.

28. Alexis MN, Gratzer WB (1978) Interaction of skeletal myosin light chains with calcium ions. Biochemistry 17: 2319-2325.

29. Morita F, Kondo S, Tomari K, Minowa O, Ikura M, et al. (1985) Calcium binding and conformation of regulatory light chains of smooth muscle myosin of scallop. J Biochem 97: 553-561.

30. Shimomura O, Johnson FH (1970) Calcium binding, quantum yield, and emitting molecule in aequorin bioluminescence. Nature 227: 1356-1357.

31. Allen DG, Blinks JR, Prendergast FG (1977) Aequorin luminescence: relation of light emission to calcium concentration--a calcium-independent component. Science 195: 996-998.

32. Klee CB, Crouch TH, Krinks MH (1979) Calcineurin: a calcium- and calmodulin-binding protein of the nervous system. Proc Natl Acad Sci U S A 76: 6270-6273.

33. Cox JA (1986) Isolation and characterization of a new Mr 18,000 protein with calcium vector properties in amphioxus muscle and identification of its endogenous target protein. J Biol Chem 261: 13173-13178.

34. Fraga H, Faria TQ, Pinto F, Almeida A, Brito RM, et al. (2010) FH8--a small EF-hand protein from Fasciola hepatica. FEBS J 277: 5072-5085.

35. Babini E, Bertini I, Borsi V, Calderone V, Hu X, et al. (2011) Structural characterization of human S100A16, a low-affinity calcium binder. J Biol Inorg Chem 16: 243-256.

36. Celic A, Petri ET, Demeler B, Ehrlich BE, Boggon TJ (2008) Domain mapping of the polycystin-2 C-terminal tail using de novo molecular modeling and biophysical analysis. J Biol Chem 283: 28305-28312.

37. Kanuru M, Samuel JJ, Balivada LM, Aradhyam GK (2009) Ion-binding properties of Calnuc, Ca^{2+} versus Mg2+--Calnuc adopts additional and unusual Ca^{2+}-binding sites upon interaction with G-protein. FEBS J 276: 2529-2546.

38. Cox JA, Tirone F, Durussel I, Firanescu C, Blouquit Y, et al. (2005) Calcium and magnesium binding to human centrin 3 and interaction with target peptides. Biochemistry 44: 840-850.

39. Durussel I, Blouquit Y, Middendorp S, Craescu CT, Cox JA (2000) Cation- and peptide-binding properties of human centrin 2. FEBS Lett 472: 208-212.

40. Miron S, Durand D, Chilom C, Perez J, Craescu CT (2011) Binding of calcium, magnesium, and target peptides to Cdc31, the centrin of yeast Saccharomyces cerevisiae. Biochemistry 50: 6409-6422.

41. Yu L, Sun C, Mendoza R, Wang J, Matayoshi ED, et al. (2007) Solution structure and calcium-binding properties of EF-hands 3 and 4 of calsenilin. Protein Sci 16: 2502-2509.

List of publications.

1. Prediction and Analysis of Canonical EF Hand Loop and Qualitative Estimation of Ca2+ Binding Affinity **Mohit Mazumder** · Narendra Padhan · Alok Bhattacharya · Samudrala Gourinath* · **PLoS ONE Apr 2014**
2. Investigations on binding pattern of kinase inhibitors with PPARγ: Molecular docking, molecular dynamic simulations and free energy calculation studies. **Mohit Mazumder**, Prija Ponnan, Umashankar Das, Samudrala Gourinath, Haseeb Khan, Jian Yang, Meena Kishore Sakharkar. **PPAR research. 2017 March**
3. N-acetyl ornithine deacetylase is a moonlighting protein and is involved in the adaptation of Entamoeba histolytica to nitrosative stress. Shahi P, Trebicz-Geffen M, Nagaraja S, Hertz R, Alterzon-Baumel S, Methling K, Lalk M, **Mazumder M**, Samudrala G, Ankri S. Nature **Scientific Reports (Sci Rep). 2016 Nov 3**
4. Crystal structure of Arabidopsis thaliana calmodulin7 and insight into its mode of DNA binding. Kumar S, **Mazumder M**, Gupta N, Chattopadhyay S, Gourinath S. **FEBS Lett. 2016 Sep**;
5. Structural insight into β-Clamp and its interaction with DNA Ligase in Helicobacter pylori. Preeti Pandey, Khaza Faisal Tarique, **Mohit Mazumder**, Syed Arif Abdul Rehman, Nilima Kumari, and samudrala Gourinath*. Nature **Scientific Reports (**Sci Rep**) July 2016**
6. Structural investigation and inhibitory response of halide on phosphoserine aminotransferase from Trichomonas vaginalis. Rohit Kumar Singh · **Mohit Mazumder** · Bhumika Sharma · Samudrala Gourinath*· **Biochimica et Biophysica Acta (BBA-General Subjects) Apr 2016** ·
7. Structure-Based Design of Inhibitors of the Crucial Cysteine Biosynthetic Pathway Enzyme O-Acetyl Serine Sulfhydrylase. **Mohit Mazumder** · Samudrala Gourinath*· **Current Topics in Medicinal Chemistry** (CTMC) Aug 2015
8. Ligand-induced conformation changes drive ATP hydrolysis and function in SMARCAL1. Meghna Gupta* · **Mohit Mazumder*** · Karthik

Dhatchinamoorthy · Macmillan Nongkhlaw · Dominic Thangminlen Haokip · Samudrala Gourinath · Sneha Sudha Komath · Rohini Muthuswami*· **FEBS Journal Jul 2015** (co. first author)

9. Crystal Structure of Calcium Binding Protein-5 from Entamoeba histolytica and Its Involvement in Initiation of Phagocytosis of Human Erythrocytes Sanjeev Kumar* · Saima Aslam* · **Mohit Mazumder** · Pradeep Dahiya · Aruna Murmu · Babu A. Manjasetty · Rana Zaidi · Alok Bhattacharya · S. Gourinath* · **PLoS Pathogens Dec 2014**
10. EhCoactosin Stabilizes Actin Filaments in the Protist Parasite Entamoeba histolytica Nitesh Kumar* · Somlata* · **Mohit Mazumder** · Priyanka Dutta · Sankar Maiti · Samudrala Gourinath* · **PLoS Pathogens Sep 2014**
11. Mutational analysis of the helicase domain of a replication initiator protein reveals critical roles of Lys 272 of B' motif and Lys 289 of β- hairpin loop in geminivirus replication. Biju George · Rajrani Ruhel · **Mohit Mazumder** · Veerendra Kumar Sharma · Swatantra Kumar Jain · Samudrala Gourinath · Supriya Chakraborty* · **Journal of General Virology Apr 2014**
12. Molecular basis of ligand recognition by OASS from E-histolytica: Insights from structural and molecular dynamics simulation studies Isha Raj · Mohit Mazumder · Samudrala Gourinath* · **Biochimica et Biophysica Acta Jun 2013** (BBA-General Subjects)
13. Crystal Structure and Mode of Helicase Binding of the C-Terminal Domain of Primase from Helicobacter pylori Syed Arif Abdul Rehman* · Vijay Verma* · **Mohit Mazumder** · Suman K Dhar* · S Gourinath* · **Journal of bacteriology Apr 2013**
14. Single residue mutation in EhSAT3 active site helps in partial regaining of feedback inhibition by cysteine. Sudhir Kumar · **Mohit Mazumder** · Sudhaker Dharavath · S Gourinath* · **PLoS ONE Feb 2013**
15. The GPI Anchor Signal Sequence Dictates the Folding and Functionality of the Als5 Adhesin from Candida albicans Mohammad Faiz Ahmad · Bhawna Yadav · Pravin Kumar · Amrita Puri · **Mohit Mazumder** · Anwar Ali · Samudrala Gourinath · Rohini Muthuswami · Sneha Sudha Komath* · **PLoS ONE Apr 2012**

16. A machine learning approach to modulate the calcium binding affinity in EF hand proteins and comparative insights into the site-specific binding affinity. **Mohit Mazumder,** Sanjeev kumar, Devbrat kunwar, Samudrala Gourinath* (manuscript under preparation).

Book Chapters:

1. Cloning, Expression and Functional Characterization of Als5: An Adhesin from Candida albicans. Sneha Sudha Komath*, Mohammad Faiz Ahmad and **Mohit Mazumder** School of Life Sciences Jawaharlal Nehru University, India

2. Structural Biology of Cysteine Biosynthetic Pathway Enzymes Isha Raj • Sudhir Kumar • **Mohit Mazumder** • S. Gourinath . **Parasitology**

International conference papers published online:

1. Upstream Sequence Finder-Tool to Find Out Upstream Element in Various Database or Genome IEEE International Advance Computing Conference IACC (2009). Vineet Jha, Mohit Mazumder, Sushanta Roy*
2. Multiple Sequence Alignment Based Upon Statistical Approach of Curve Fitting. Vineet Jha, Mohit Mazumder*, Hrishikesh Bhuyan, Ashwani Jha DOI: 10.1007/978-3-642-11164-8_30 · Source: DBLP Conference: Pattern Recognition and Machine Intelligence, Third International Conference, PReMI 2009, New Delhi, India, December 16-20, 2009 Proceedings

OPEN ACCESS Freely available online

PLOS | ONE

Prediction and Analysis of Canonical EF Hand Loop and Qualitative Estimation of Ca^{2+} Binding Affinity

Mohit Mazumder[1], Narendra Padhan[1,2], Alok Bhattacharya[1,3], Samudrala Gourinath[1]*

1 School of Life Sciences, Jawaharlal Nehru University, New Delhi, India, **2** Department of Immunology, Genetics, and Pathology, Rudbeck Laboratory, Uppsala University, Uppsala, Sweden, **3** School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi, India

Abstract

The diversity of functions carried out by EF hand-containing calcium-binding proteins is due to various interactions made by these proteins as well as the range of affinity levels for Ca^{2+} displayed by them. However, accurate methods are not available for prediction of binding affinities. Here, amino acid patterns of canonical EF hand sequences obtained from available crystal structures were used to develop a classifier that distinguishes Ca^{2+}-binding loops and non Ca^{2+}-binding regions with 100% accuracy. To investigate further, we performed a proteome-wide prediction for *E. histolytica*, and classified known EF-hand proteins. We compared our results with published methods on the E. *histolytica* proteome scan, and demonstrated our method to be more specific and accurate for predicting potential canonical Ca^{2+}-binding loops. Furthermore, we annotated canonical EF-hand motifs and classified them based on their Ca^{2+}-binding affinities using support vector machines. Using a novel method generated from position-specific scoring metrics and then tested against three different experimentally derived EF-hand-motif datasets, predictions of Ca^{2+}-binding affinities were between 87 and 90% accurate. Our results show that the tool described here is capable of predicting Ca^{2+}-binding affinity constants of EF-hand proteins. The web server is freely available at http://202.41.10.46/calb/index.html.

Citation: Mazumder M, Padhan N, Bhattacharya A, Gourinath S (2014) Prediction and Analysis of Canonical EF Hand Loop and Qualitative Estimation of Ca^{2+} Binding Affinity. PLoS ONE 9(4): e96202. doi:10.1371/journal.pone.0096202

Editor: Rajagopal Subramanyam, University of Hyderabad, India

Received February 25, 2014; **Accepted** April 4, 2014; **Published** April 23, 2014

Funding: This study was supported by Innovative Young Biotechnology Award, Department of Biotechnology, and Govt. of India (http://www.dbtindia.gov.in/index.asp). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: samudralag@yahoo.com

Introduction

Calcium signaling plays a major role in controlling most biological systems and many cellular functions, such as fertilization, motility, cell differentiation, proliferation and apoptosis, which are directly or indirectly regulated by Ca^{2+} [1–3]. In eukaryotes, there are elaborate mechanisms that are involved in maintaining Ca^{2+} homeostasis [4]. A defect in any of the components of the Ca^{2+} homeostasis/signaling system may have disastrous consequences including cell death. Recently many Ca^{2+}-binding proteins have also been identified in bacteria and viruses, raising the possibility that the prokaryotes may also have a Ca^{2+} regulatory system, particularly in relation to host-pathogen interactions [5,6].

Ca^{2+} is bound by a variety of proteins that are capable of binding with different affinities [7–9]. Such calcium binding proteins (CaBPs) can be classified into two categories, Ca^{2+} sensors and buffers. The major function of the first category of CaBPs is to sense the level of free intracellular $Ca^{2+}$and then to activate a suitable signaling pathway [10].

In general, CaBPs contain two well-defined Ca^{2+}-binding motifs: the EF hand and C2 domains [11]. The EF-hand motif is the most frequently occurring Ca^{2+}-binding motif in eukaryotic systems [12]. There are more than 66 subfamilies [13] of EF–hand proteins and 3000 EF-hand related entries in the NCBI Data Bank [14]. An EF hand is composed of a typical helix-loop-helix structural unit. This group is the largest and includes well-known members, such as calmodulin, troponin C and S100B. These proteins typically undergo a calcium-dependent conformational change which opens a target binding site [13]. Proteins, such as calbindin D9k do not undergo calcium-dependent conformational changes [15–17].

EF-hand motifs are divided into two major structural groups: the canonical EF-hands as seen in calmodulin (CaM) and the prokaryotic CaM-like protein calerythrin, and the pseudo EF hands exclusively found in the N-termini of S100 and S100-like proteins [18]. In either structural group, a pair of EF-hand motifs or pseudo EF-hand motifs forms a structural domain and is the minimum requirement for Ca^{2+}-dependent activation. In general, one of the EF-hand motifs has a higher Ca^{2+}-binding affinity than the other. The canonical Ca^{2+}-binding loop is characterized by a sequence of 12 amino acid residues. In an EF-hand loop the calcium ion is coordinated in a pentagonal bipyramidal configuration. The six residues involved in the binding are in positions 1, 3, 5, 7, 9 and 12; these residues are denoted by X, Y, Z, -Y, -X and -Z.

In general, affinity constants of EF-hand domains for Ca^{2+} vary from micromolar to millimolar, reflecting the diversity of functions carried out by these proteins in a range of Ca^{2+} concentrations. There is an increase in stability and change in conformation upon binding Ca^{2+}. Several residues found in an EF-hand loop are highly conserved and contribute to the stabilization and proper folding of the binding site. Factors such as biological environment as well as the binding sequence have been shown to contribute to the calcium-binding affinity of these proteins [18–21].

A number of algorithms have been developed to computationally identify EF hand-containing CaBPs and Ca^{2+}-binding regions, including statistical, machine learning and pattern search approaches [22–24]. Recently, Franke et al. (2010) [24]proposed a method to estimate Ca^{2+}-binding affinity based on free energy calculations using crystal structures of CaBPs. However, this method has limited use due to unavailability of crystal structures in complex with calcium for large number of CaBPs. Moreover, no suitable method is available for the prediction of Ca^{2+}-binding affinity from primary sequence information. There was an early attempt by Boguta et al (1988) [25] to estimate the binding affinity of calcium for troponin C (TnC) superfamily proteins based on the prediction of secondary structures. The results were convincing for some proteins which follow a typical TnC pattern [25] but not for any other protein family. Since it is not always possible to experimentally determine Ca^{2+}-binding properties of EF hand-containing calcium-binding proteins, it is necessary to be able to predict this property from primary sequence. In this report we describe a method for computational prediction of Ca^{2+}-binding loops and their affinities for $Ca^{2+}$from amino acid sequences. This paper describes approaches to find a better correlation of sequence to binding affinities in order to predict the sequence to function (Ka) relationship.The results show that the tool (CAL-EF-AFi) described here is accurate and provides useful information about Ca^{2+}-binding properties to experimental biologists for both characterized and uncharacterized proteins.

Results

A few experimental methods based on biophysical techniques, such as Isothermal titration calorimetry (ITC) surface plasmon resonance (SPR) & fluorescence [26] are available for determination of Ca^{2+}-binding parameters. However, these are expensive and time consuming. To the best of our knowledge, no prediction method has been developed so far that can be used to estimate Ca^{2+}-binding properties of a protein from primary sequence. Therefore, a comprehensive study was carried out first to identify Ca^{2+}-binding EF loops and then their Ca^{2+}-binding affinities. In this study, we have constructed two support vector machines (SVM), one for prediction of loop regions and the other for estimation of binding affinity.

Position-specific scoring matrix

After obtaining position-specific scoring matrix (PSSM) scores using equations (1) and (2) (described in Methods) for all the sequences obtained from the literature, we calculated the correlation coefficient between the experimental affinity constants (Ka) and PSSM to be 0.61 (Figure S1 in File S1). While this correlation is clearly positive, it was not possible to classify the affinity of all the sequences solely using PSSM scores. Therefore, a systematic attempt was made to first predict the presence of canonical EF-hand loops from amino acid sequence and then estimate the binding affinities qualitatively based on evolutionary information using SVMs.

Amino acid composition distinguishes Ca^{2+}-binding and non-binding regions

A statistical analysis was carried out to determine which amino acids are found unusually frequently in EF hand-motif sequences using the entire PFAM EF-hand database. Glycine, glutamic acid, asparagine, and especially aspartate have been determined to occur more frequently in Ca^{2+}-binding loop regions than in non-binding regions at a 99.9% confidence level. Alanine, phenylalanine, leucine, and especially methionine are overrepresented in non-binding regions (Figure 1). The relative frequency of amino acids at each position is listed in Table S1 in File S1. The analysis suggests that EF-hand Ca^{2+}-binding loops have a specific amino acid composition, and that it is possible to identify these loops from the primary sequence.

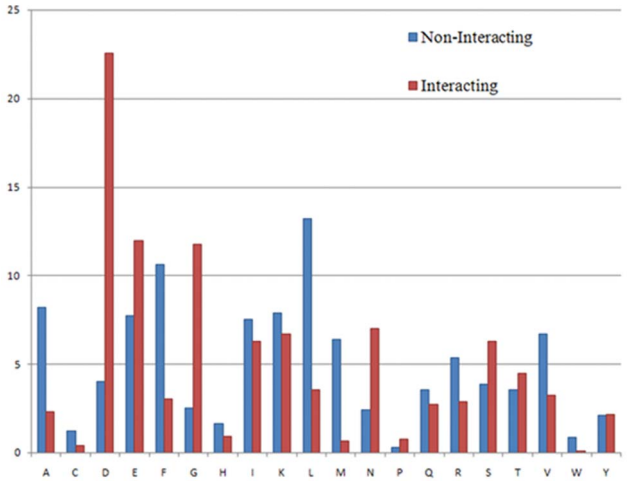


Figure 1. Amino acid composition of the 12-mer long Ca^{2+}-binding region (''Interacting'') and the non-binding region (''Non-Interacting'') of EF-hand proteins.
doi:10.1371/journal.pone.0096202.g001

Experimental determination of Ca^{2+}-binding properties of EhCaBPs

In order to validate the theoretical predictions, experiments were carried out to determine qualitative and quantitative aspects of the affinity of some EhCaBPs for Ca^{2+}. Ca^{2+}-binding properties of these proteins were tested by $^{45}Ca^{2+}$ overlay assay on western blotted pure recombinant EhCaBP1, 3, 5, 6, and 7 proteins. All of these proteins were found to bind $^{45}Ca^{2+}$ as observed by autoradiography (data not shown). ITC was used to determine the molar stoichiometry of the binding of the cations to these EhCaBPs, as well as the binding constants and associated thermodynamic parameters (Table 1). The sequences and binding affinities of these proteins were used in the validation dataset (D7) for validation of the classifier's efficiency on experimental data. The raw data obtained after ITC experiments are provided in the Figure S2 in File S1.

SVM models predict the presence of EF loop regions

Two different models were generated using both binary pattern and amino acid composition (AAC) for loop identification. Both AAC and binary pattern were calculated, and used as input for classification of Ca^{2+}-binding EF-hand loops and non-Ca^{2+}-binding 12-mers in EF-hand proteins using SVM. The models were generated by using different types of kernels, such as polynomial, radial basis function (RBF), and linear. The performance of each kernel function was evaluated by five-fold cross validation. During model generation, the RBF kernel showed the best results.

The RBF kernel function using binary and AAC standalone features most accurately predicted the presence of EF-loop regions. An accuracy of 100% was achieved with D1 and D2. The remarkable performance of binary and AAC is due to the high conservation of sequence and structure among EF-hand loops that have been used in this study. Normally, the default threshold

Table 1. Summary of macroscopic binding constants and thermodynamic parameters obtained from the ITC studies of Ca^{2+}-binding isotherm of EhCaBPs at 25°C.

Ligand	Titrand	No of experimental Ca^{2+}-binding sites (n)	KA (M-1)	Kd	ΔH (cal/mol)	ΔS (cal/mol)	ΔG (kcal/mol)
Ca^{2+}	**EhCaBP1**	4	$K1 = 5.25\times10^{3}\pm4.0\times10^{2}$	130.72 μM	-1860 ± 0	10.8	-4.84
			$K2 = 1.41\times10^{4}\pm9.5\times10^{2}$		$2.3\times10^{5}\pm0$	790	-4.6×10^{2}
			$K3 = 5.10\times10^{5}\pm2.8\times10^{4}$		$2.4\times10^{5}\pm1.82\times10^{3}$	-780	-7.56
			$K4 = 1.55\times10^{6}\pm7.3\times10^{4}$		$-7981\pm1.86\times10^{3}$	1.56	-8.44
	EhCaBP3	2	$K1 = 4.00\times10^{6}\pm5.3\times10^{5}$	1.85 μM	$-1.605\times10^{4}\pm86.6$	-23.6	-9.0
			$K2 = 7.28\times10^{4}\pm5.3\times10^{3}$		-7573 ± 10^{4}	-3.16	-6.63
	EhCaBP5	2	$K = 1.18\times10^{7}\pm1.47\times10^{6}$	85 nM	$-1.84\times10^{4}\pm61.79$	-29.4	-9.64
	EhCaBP6	2	$K1 = 1.07\times10^{5}\pm1.1\times10^{4}$	46 μM	702 ± 17.6	25.4	-6.86
			$K2 = 4.44\times10^{3}\pm1.1\times10^{2}$		5244 ± 45.9	34.3	-4.97
	EhCaBP7	2	$K1 = 1.04\times10^{6}\pm2.5\times10^{5}$	3.12 μM	-1807 ± 96.5	21.5	-8.2
			$K2 = 9.86\times10^{4}\pm6.8\times10^{3}$		-5413 ± 96.5	4.69	-6.81

doi:10.1371/journal.pone.0096202.t001

value (0) was used for the SVM classifier to discriminate between Ca^{2+}-binding EF-hand loops and non-Ca^{2+}-binding 12-mers in EF-hand proteins. The sites with a prediction score close to 1 are most likely to be an EF-hand calcium-binding loop region. All performance measures and the learning parameters for the RBF kernel are listed in Table 2.

Accessibility and hydrophilic (AC&HC)-based classifier provides the best estimation of binding affinity

Various SVM models using a combination of features were developed to estimate the affinity of Ca^{2+} for the EF-hand loop. The predictions of binding constants were not as accurate as the predictions of EF-hand loops due to the limited availability of experimental data on binding constants and the high level of diversity in amino acid sequence with relation to binding affinity. In this study, we have developed a position-specific scoring matrix for EF-hand loop regions and scored (equation [1] and [2]) the sequences from the annotated data set using Perl scripts developed in-house. Based on the PSSM scores, we classified high (D3) and low (D4) binding groups for the 12-mer region to train the classifier. The binding constants, obtained from the literature (Table S2 in File S1), and data obtained from ITC studies of EhCaBPs were used as the test dataset and validation dataset (Table S3 in File S1) respectively. Since it is generally believed that different physico-chemical properties contribute to the structure and function of protein sequences, these properties should also contribute to Ca^{2+}-binding affinity. Therefore, we have developed several SVM models (data not shown) to achieve better accuracy using combinations of several amino acid features, and have obtained the different physico-chemical properties using the amino acid index database (http://www.genome.jp/aaindex/).Only the best performing models are discussed here.

For the 24-dimension input vectors consisting of accessibility (AC) and charge (CC), the values of sensitivity, specificity and accuracy were 90.97, 87.10, 90.30 and 90.91, 75.00, 84.21 for training and test datasets respectively. We were also able to achieve a Matthews's correlation coefficient (MCC) of 0.78 for the training datasets (D3 & D4) and 0.67 for the test (D5) dataset.

The classifier consisting of concatenated features of accessibility (AC) and hydrophilic (HC) scores showed the best performance when tested on the training and the test datasets, achieving an MCC of 0.87 and 0.81 and an accuracy of 94.78 and 89.47 for D3–D4 and D5 datasets, respectively. The superior performance of this classifier compared to other hybrid models is also indicated by its values for sensitivity and specificity of 95.83 and 91.00 respectively for the training dataset, and 81.82 and 100.0 respectively for the test dataset.

Several other hybrid models (AC&CC, AC&HC&HYC, AC&HYC&CC and AC&HYC) were also generated with amino acid features-based classifiers; however their performances were not better than the AC&HC-based classifier. The list of figures of merit of all the classifiers used can be found in Tables 3 and 4.

The quality of the performance of the AC&HC-based classifier is also indicated by receiver operating characteristic (ROC) plots, which we computed for all the models discussed in this study. ROC is commonly used to evaluate the discrimination ability of a classifier. If the area under the ROC curve is larger, it means the classifier has better discrimination ability. We were able to achieve an AUC of 0.97 with the training dataset and 0.903 with the experimental datasets (D5 & D7) using the AC&HC-based classifier (Figure S3 in File S1). A schematic representation for the data input, algorithm implementation and experimental strategy overview is shown in Figure S4 in File S1.

Table 2. The Performance of SVM Models with different learning parameters on D1 and D2 dataset.

Features	C	g	SN	SP	ACC	MCC
Binary	8	0.008	100	100	100	1
AA	0.125	0.008	100	100	100	1

Using binary patterns and AA (amino acid) composition [γ **(g)** (in RBF kernel), c: parameter for trade-off between training error & margin] where SN–sensitivity, SP–specificity, ACC-accuracy, MCC–Matthews Correlation Coefficient.
doi:10.1371/journal.pone.0096202.t002

Prediction of Ca^{2+}binding of an independent dataset

After obtaining the best performing model, it was important to evaluate the performance of this classifier on a dataset that has not been used for training and testing. In order to check the unbiased prediction efficiency of the model, in addition to the test dataset, an independent dataset (D6) with 35 unique troponin C superfamily binding sites (Boguta et al 1988) and 15 unique sites (Table S4 in File S1) were tested using our classifier. The classifier predicted 21 high binders (true positives), 19 low binders (true negatives), and 10 high binders (false negatives) that were predicted as low binding sites. When using the diverse datasets and binding affinities obtained from different researchers working under different experimental conditions, the overall accuracy achieved was 80.0%.

The validation dataset

The performance of AC&HC-based classier was even better when tested on the experimentally obtained binding affinities from EhCaBPs. We achieved an accuracy of 90.91 and MCC of 0.83. The performances of other classifiers for the validation dataset D7 are listed in table 5.

E. histolytica proteome analysis: Computational prediction of Ca^{2+}-binding properties of EhCaBPs

In this section, we used 'CAL-EF-AFi' to scan the *E. histolytica* proteome in order to predict all Ca^{2+}-binding canonical EF-hand loops in this organism. A previous computational study [27] showed that there are 27 CaBPs containing EF-hand motifs present in *E. histolytica*. Our scanning results picked all the known canonical EF hands with more than one EF-hand loop region. Apart from the sequences used in the test dataset (Ehcabp1, 3, 5–7); we also predicted the relative affinities of other EhCaBPs (8–27). In total, we predicted 36 Ca^{2+}-binding sites (Table S5 in File S1) out of which 24 were predicted to be low-affinity sequences and the remaining 12 sites were predicted to have high affinity for Ca^{2+}.

Comparison with existing methods

The performance of the classifier was compared with PFAM based HMM profile search and Calpred [28] on the *E. histolytica* proteome. In light of earlier bioinformatics studies by Bhattacharya et al. and availability of *E.histolytica* strain HM-1: IMSS for wet lab experiments, we chose the *E.histolytica* proteome for comparison. Although this is not a benchmark dataset, it was important to validate our classifier's accuracy to find EF-hand containing Ca^{2+}-bindingsites in large databases and proteomes. A total of 41 EF-hand protein sequences were predicted using the pattern search method whereas CAL-EF-AFi found 58 probable sequences with 153 binding loops.

Based on the results obtained by PFAM pattern search, few of the predictions with high threshold values (Table S6 in File S1) appear to be false positives. Note that the tertiary structures of all these proteins have not been determined yet, but lacks the number of amino acids required to form a typical EF hand structural motif. Similarly we scanned EhCaBPs with Calpred (using all the modules available), which identified EF-hand proteins but predicted false positives; all the residues in the full-length protein sequence were predicted as calcium binding (site). To investigate further we used sequences with known structures (D1 & D2) in Calpred and found similar false-positive predictions here as well. A thorough analysis (Table S6 in File S1) of the results from different methods for the identification of EF-hand Ca^{2+}-binding sites suggests that the method proposed here to be most suitable for prediction of Ca^{2+}-binding sites and relative affinity constants and is also useful for whole proteome scans.

Availability

CAL-EF-AFi is available at http://202.41.10.46/calb/index.html and all the datasets used in the study as well as the proteome scan results are available at http://202.41.10.46/calb/dataset.html.

Discussion

In the current era of high-throughput next generation sequencing, where a large amount of genomic data is generated each day, prediction of gene functions and detailed annotation have become key aspects of computational genomics. The focus of this study is to annotate Ca^{2+}-binding EF hand motif-containing proteins and further classify these on the basis of their Ca^{2+}-binding affinities.

Different Ca^{2+}-binding proteins display different levels of affinities for Ca^{2+}. The functions of these proteins in general depend on their affinity constants for Ca^{2+}. Ca^{2+}-sensor proteins such as calmodulin (CaM) display higher Ca^{2+} -binding affinities for their C-terminal domains than for their N-terminal domains [29]. Ca^{2+}-buffer proteins, such as parvalbumin have high binding affinity [30] and there is little or no change in their conformation upon binding Ca^{2+}. Hence it is possible to predict the probable function of the proteins from Ca^{2+}-binding properties.

Many computational methods have been developed ever since identification of the first EF-hand domain as an approach for prediction of Ca^{2+}-binding sites. These methods were based on similarity search, energy based calculations, Bayesian statistical methods, machine learning approaches and graph theory [22,31–33], where the input is either a primary amino acid sequence or a three-dimensional structure. A comparison of CAL-EF-AFi with the existing methods for identifying Ca^{2+}-binding sites is not suitable due to the dissimilarity in the prediction methods, input type and the datasets. One of the recently published machine learning approaches [28] to identify the calcium-binding region showed poor performance when compared with CAL-EF-AFi using a dataset of experimentally determined values. Some of the other methods, such as CaPS uses pattern search where EF-hand motif and Ca^{2+}-binding loops are predicted on the basis of

Table 3. The Performance of SVM Models on PSSM based training dataset D3 & D4.

Features	C	g	SN	SP	ACC	MCC	AUC/ROC
AC&CC	32768	0	90.97	87.1	90.30	0.78	0.94
AC&HC	***8***	***0.03***	***95.83***	***91.0***	***94.78***	***0.87***	**0.97**
AC&HC&HYC	2	0.13	94.44	91.0	94.78	0.86	0.97
AC&HYC&CC	2048	0	91.67	90.32	91.42	0.82	0.96
AC&HYC	2048	0	91.67	88.7	91.04	0.8	0.95

The Performance of SVM Models on PSSM based training dataset D3 & D4 with different learning parameters on various hybrid models [γ (g) (in RBF kernel), c: parameter for trade-off between training error & margin] where SN–sensitivity, SP–specificity, ACC-accuracy, MCC–Matthews Correlation Coefficient, AUC/ROC-Area under curve/ Receiver Operating Curve.

doi:10.1371/journal.pone.0096202.t003

patterns generated using a Hidden Markov Model based on multiple sequence alignment of known EF-hand proteins. None of these methods, however, were able to predict the binding affinity of the identified Ca^{2+}-binding motifs. We have trained the classifier using the sequences of EF hand motif binding and non-binding regions so that it could identify the Ca^{2+}-binding region in the EF-hand motif.

The performance of the classifier was also tested by analysing the complete proteome of *E. histolytica*. Based on the scan results we found all of the reported Ca^{2+}-binding proteins, and also identified new probable Ca^{2+}-binding sites. Our tool appeared to give better results in terms of identification of CaBPs as it identified more proteins including all known CaBPs. Other methods, such as PFAM-based HMM profile search and Calpred showed a significant number of false predictions. Our results, using all of the sequences in the test (D5) affinity estimation data set, suggest that the PSSM scores and experimental binding affinities are broadly correlated. In our study, we have classified proteins on the basis of relative binding affinity for Ca^{2+} in a semi-quantitative manner. There are a number of reasons that a precise quantitative analysis is still intractable. For one, a 12-mer motif alone does not determine the affinity since there may be contributions from other parts of the protein. Also, there is a cooperative involvement of more than one EF-hand loop in the binding of Ca^{2+}. This may be particularly important as a pair of EF hands occur together [14]. Two EF-hand motifs in a pair (with very few exceptions) are related by an approximate two-fold rotational axis, forming a hydrophobic cavity opening which is likely to influence the binding affinity. Since these properties are difficult to factor in a model, our efforts are limited to classification of high and low binders rather than predicting precise binding affinities.

Our initial datasets contained 19 binding sites with experimental binding affinity data. In order to circumvent the problems associated with limited data, we have generated training datasets based on the evolutionary information (PSSM) scores. A similar approach, where artificial datasets have been used in SVM, has been successful in greatly improving predictions [34,35]. In these studies, researchers have mainly generated negative datasets artificially for SVM classification. Our test data set with 19 sequences, independent dataset with 50 sequences and the validation data set with 11 sequences representing experimentally determined affinity data have shown extremely good results.

The results from the test and validation datasets, which includes relative affinities of several EF-hand proteins, suggest that our proposed model based on the PSSM method for estimation of binding affinity can help researchers to predict site-specific binding affinity. Experimental determination of such binding affinity is a limiting factor in Ca^{2+}-binding proteins because of the expense involved and time required carrying out the experiments. As mentioned above, the successful performance of the model with regards to prediction and estimation is attributed to the accurate training of the classifier on a small number of training examples and the use of PSSM generated datasets.

CAL-EF-AFi can therefore be used to accurately and precisely scan proteomes of organisms for potential Ca^{2+}-binding sites of EF-hand proteins and estimate their probable relative binding affinities. Given the success of our classifier on the *E. histolytica* proteome scan, we expect its wider use in analysing proteomes of other organisms.

In conclusion, we have developed a unique method, CAL-EF-AFi for identification and estimation of Ca^{2+}-binding sites and relative affinity. The program requires only the protein sequence for the prediction without prior knowledge of structural or biochemical information. The results predicted by the theoretical

Table 4. The Performance of SVM Models on test dataset D5.

Features	SN	SP	ACC	MCC
AC&CC	90.91	75.00	84.21	0.67
AC&HC	**81.82**	**100**	**89.47**	**0.81**
AC&HC&HYC	72.73	87.50	78.95	0.6
AC&HYC&CC	90.91	75.00	84.21	0.67
AC&HYC	90.91	75.00	84.21	0.67

The Performance of SVM Models on test dataset D5 (experimental binding affinities obtained from literature) with different learning parameters.
doi:10.1371/journal.pone.0096202.t004

model were validated by experimental studies. Variation from the EF-hand consensus sequence can be used to predict qualitative Ca^{2+}-binding features. However, this may not be sufficient to understand the overall characteristics of CaBPs. The EF-hand motifs assemble to form a lobe (one partner affects the binding affinity of the other) and the Mg^{2+} affinities are not considered in this work due to limitation of experimental data available to date. Future plans include developing an even better algorithm with more information available from the literature. We hope that an increase in the availability of experimental data will help generate a more robust model.

Material and Methods

Expression, Purification and Preparation of Metal-free Protein Solutions

Five different EhCaBPs (EhCaBP1, 3, 5, 6, and 7) were overexpressed and purified as described earlier [36,37]. In order to obtain accurate measurements of Ca^{2+}-binding energetics, it was essential to have the protein in its apo-form with no contamination of Ca^{2+} in the buffers. Hence, all of the buffers used for isothermal titration calorimetry (ITC) were decalcified using Chelex 100 resin (Bio-Rad). Decalcified ITC buffer (100 mMNaCl and 50 mM Tris-Cl, pH 7.0) was prepared by treatment with Chelex 100 resin (Bio-Rad). Each protein solution was treated with 5 mM EGTA and 2 mM EDTA to remove Ca^{2+} and Mg^{2+}. The EDTA/EGTA bound to metal ions were removed from protein solution using Amicon ultra centrifugal filter devices (Millipore), through extensive buffer exchange (decalcified). Before the ITC experiment, the sample cell and injection syringe of the ITC machine (Microcal Inc.) were extensively cleaned using the decalcified buffer.

Isothermal Titration Calorimetry (ITC)

All ITC experiments were performed on a MicroCal VP-ITC microcalorimeter at 25 C. Samples were decalcified, centrifuged, and degassed prior to titration. A typical titration consisted of injecting 2-µl aliquots of 10–20 mM $CaCl_2$ solution (diluted from 1 M standard $CaCl_2$ solution supplied by Sigma-Aldrich Chemicals) into 100–200 µM protein solution after every 3 min to ensure that the titration peak returned to the baseline prior to the next injection. A total of 70 injections were carried out. Aliquots of concentrated ligand solution were injected into the buffer solution (without the protein) in a separate ITC run, to subtract the heat of dilution. Two sets of titrations were carried out for each protein: (i) apo-EhCaBP in 50 mM Tris-Cl, pH 7.0 and 100 mMNaCl and (ii) holo-EhCaBP in 50 mM Tris-Cl, pH 7.0 and 100 mMNaCl. The ITC data were analysed using the software ORIGIN (supplied with Omega Microcalorimeter). The amount of heat released per addition of the titrant was fitted to the best least squares model as given by Wiseman et al. (1989). For each titration, the stoichiometry (n), association constant (Ka), and enthalpy change (ΔH) were obtained directly from the ITC data, and the changes in Gibbs free energy (ΔG), and entropy (ΔS), as well as the overall binding affinity or dissociation constant (Kd) were calculated according to Equations a, b, and c.

$$\Delta G = RT \ln Ka \quad \text{(a)}$$

$$\Delta G = \Delta H - T\Delta S \quad \text{(b)}$$

Table 5. The Performance of SVM Models on validation dataset with experimentally derived binding affinity from EhCaBPs (D7).

Features	SN	SP	ACC	MCC
AC&CC	83.33	60	72.73	0.45
AC&HC	**100**	**80**	**90.91**	**0.83**
AC&HC&HYC	83.33	80	81.82	0.63
AC&HYC&CC	83.33	60	72.73	0.45
AC&HYC	66.67	60	63.64	0.27

The Performance of SVM Models on validation dataset with experimentally derived binding affinity from EhCaBPs (D7)with different learning parameters on various hybrid models [γ (g) (in RBF kernel), c: parameter for trade-off between training error & margin] where SN–sensitivity, SP–specificity, ACC-accuracy, MCC–Matthews Correlation Coefficient, AUC/ROC-Area under curve/ Receiver Operating Curve.
doi:10.1371/journal.pone.0096202.t005

$$Kd = 1/Ka \text{ or } Kd = 1/\sqrt{K1K2K3.....} \quad (c)$$

Dataset for EF loop predictions

To predict the presence of EF-hand loops and estimate their affinities for Ca^{2+}, the calcium-binding amino acid sequence pattern at PROSITE [38](http://prosite.expasy.org/PDOC00018) was used to retrieve sequences of the EF-hand family. In total 1379 different sequences were obtained. To further validate the reviewed sequences we used structures of proteins co-crystallized with calcium from the Protein Data Bank [39] (PDB, http://www.rcsb.org/pdb/). In total 1261 chains with EF-hand motifs were found. Once these sequences were downloaded, CD-HIT [40] was used to remove redundant sequences having more than 60% similarity. The PDB IDs are included in the supplementary data in File S1 (Tables S7–S10 in File S1) along with the sequences retrieved. We chose a relatively high because the aim of the study was to identify the binding loop, which is a highly conserved 12-residue sequence. With less than a 60% threshold, the numbers of sequences available for classification were not sufficient. The sequence classifications were also carried out using thresholds of 90%, 70%, 60%, 50% of CD-HIT data is also shown in Table S11 in File S1. Finally a dataset of 100 12-mer calcium-binding loop sequences for the positive training dataset (D1) was generated. Similarly a negative training dataset was built with 141 (D2) 12-mer sequences extracted from non-binding regions of EF-hand proteins.

Dataset for binding affinity predictions

For the estimation of binding affinity, a novel method was developed on the basis of PSSM score pattern in which calcium-binding loops were classified into two groups. Based on the correlation obtained between the PSSM scores and experimental binding affinity (Figure S1 in File S1) a positive dataset with high PSSM scores (D3) (>5) consisting of 144 12-mer sequences and a negative dataset (D4) with low PSSM scores (<5) containing 124 sequences were generated using the sequences obtained from PROSITE [38].

To test the proposed model based on PSSM scores we used 19 EF loop sequences for which binding affinities were known from the literature (Table S2 in File S1) as Test dataset (D5). To evaluate the performance of this classifier on a dataset that has not been used for training and testing, an independent dataset (D6) of binding affinity observations was obtained from Boguta et al (1988) [25]and recently published literature. After removing redundant EF-loop sequences, 50 unique sequences were obtained from recently published data and the Ka values listed in Boguta et al (1988) [25].Furthermore, to check the performance and reliability of the classifier, we chose to perform ITC experiments on available EhCaBPs, to test our predictions on the datasets obtained from literature. We were able to obtain Ka values of EhCaBP1, 3, 5, 6, and 7; in total we listed affinities for 11 sites used here as a validation set (D7). The details of ITC experiments and results are also provided in supplementary datasets in File S1 as D5, D6 and D7 with their experimental binding affinities classified on the basis of a thorough review of published papers that reported the binding constants. The classification details with supportive binding constants are listed under "Author's Note" in Tables S2–S4 in File S1.

Statistical Analysis

The expected (Exp) frequencies of amino acid residues were calculated from the average residue usage from the 1379 different sequences obtained from PROSITE [38]. The expected frequency for an amino acid residue of type A at position i will be Exp = *(NA/N) M*, where *NA* = total number of amino acid residues of type A in the analysed set of sequences, excluding position *i*, *N* = total number of all amino acid residues in the analysed set of sequences, excluding position *i*, and *M* = total number of sequences, i.e., the sum of *i*th positions in the analysed set of sequences. The expected frequencies for residues were calculated similarly. For each amino acid residue at a given position, the deviation of the observed (Obs) values from the Exp values was estimated by the χ^2 criterion according to the formula $(\text{Obs} - \text{Exp})^2/\text{Exp}$. For each residue or codon, the χ^2 value was estimated separately with one degree of freedom. The sums of all 20 (61) χ^2 values for each residue (codon) at the given position gave the total deviation for the given position with 19 (60) degrees of freedom. To evaluate the range of differences between the C-terminal regions and the neighbouring fragments, a pairwise comparison between them was performed. For this purpose, each position in the sequence was treated as a set containing 20 groups of data and the difference between them was calculated by the χ^2 criterion using the following formula:

$$\sum_{i=1}^{K}[(m_i/M - n_i/N)^2 MN/(m_i + n_i)]$$

where m_i and $n_i$are frequencies of amino acid residues in the two positions of the sequence under comparison, M and N are total numbers of amino acid residues in the compared positions, and K is equal to 20 because each position may be occupied by any of 20 different amino acids. At a significance level <0.001, Obs was considered to be different from Exp if the $\chi 2$ exceeded 10.8, 43.8 and 99.6 for one, 19 and 60 degrees of freedom, respectively.

Generation of a position-specific scoring matrix

In this study, a simple position-specific scoring matrix (PSSM) was generated from the amino acid composition (AAC) of the calcium-binding loops in canonical EF hands. The standard amino acid frequencies, which show how often each residue was found in each site in the binding loop, was taken from Marsden et al., 1990 [41]. In this matrix, every column can be interpreted as a discrete probability distribution of the amino acid residues at that position and the values in the matrix can be inferred as probabilities of a given amino acid occurring at a given position. Therefore, for a sequence of length m, the product of the relative frequencies from the matrix corresponding to each amino acid in each position of the sequence is the probability of discovering such a sequence in the EF-hand loop. We generated two different scoring matrices, one with simple relative frequency of amino acids and the other with log likelihood frequency for the position-specific scoring matrix [42–44]. The log ratio matrix was generated using equation 1 and 2.

$$Sij = q + bPi/n + b \quad (1)$$

$$Msij = \log(Sij/Pi) \quad (2)$$

Where Sij is the probability of amino acid i at position j in matrix S, q is the observed counts of amino acid type i at position j, Pi is the probability of amino acid type i, b is the pseudo count which is considered here as square root of the total number of training sequences and n is the number of training sequences. In equation (2) Msij represents the foreground model (representing true homology) and Pi is the background model (chance that a match occurs at random). The background probability or the chance of amino acid match occurrence at random was calculated using the BLOSUM62 substitution matrix [45].

Support Vector Machine training for classification

SVM is a machine learning tool that is being extensively used for classification and optimization of complex problems. It is particularly attractive to biological sequence analysis due to its ability to handle noise, large datasets, large input spaces and high variability [46,47]. In this study all of the SVM models have been developed using libSVM [48]. Parameter selection was carried out using grid search so that the classifier can accurately predict unknown test data from the model. In the radial basis function (RBF) kernel, there are two parameters, C and g, but it is not known *a priori* what values of these two parameters are best for a given problem [48]. To obtain the best parameters, a grid search was carried out using cross validation. A Perl script was written in-house to check combinations of features in an iterative manner using CUDA based libSVM [49]. A descriptive flowchart of the feature selection algorithm is provided in Figure S4 in File S1.

Five-fold cross-validation

A standard five-fold cross-validation technique was used to evaluate the performance of models, where the data set was randomly divided into five sets. The classifier was trained on four sets and the performance was assessed on the remaining fifth set. The process was repeated five times so that each set could be used once for testing. Finally, the average of the five sets was calculated as the measure of the performance of the classifier.

SVM model using binary and amino acid composition features

In this method, a Perl program was written to generate a window with 12 amino acids for negative and positive patterns. These sequence patterns were converted into binary patterns, where a pattern of length L was represented by a vector of dimension L×21 and each amino acid in that pattern was represented by a 21-feature vector (e.g. Asp by 1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,X) containing 20 amino acids and a dummy X. Each sequence of twelve amino acids was represented by 252 input vectors during model generation. The binary profile has been used in a number of existing methods [50,51]. The second feature used was AAC with an input vector of 20X12 dimensions. AAC is the fractional occurrence of each amino acid in the protein sequence.

$$Fi = Total\ number\ of\ Amino\ acid = Length\ of\ the\ protein$$

Where i can be any of the amino acids.

Feature extraction and model generation for binding affinity estimation

It has been observed in different studies [52,53] that SVM performs well when combinations of two or more features are used as input vectors. Hence, hybrid models have been developed using one or more combinations of features. After testing combination of features using CUDA-based libSVM [49] the best performing features were used for developing various SVM models. Feature selection was carried out by scanning amino acid indices and by performing 5-fold cross validation using the in-house CUDA script. The four best performing amino acid properties used further for analysis were net charge [54](CC), hydrophobicity [55] (HYC), hydrophilicity [56] (HC) and accessibility [57] (AC) which were thus used for further analysis. Only the better performing models(AC&CC, AC&HC, AC&HYC, AC&HC&HYC, and AC&HYC&CC), which use combinations of the four best performing amino acid properties, are discussed in this study.

Classifier performance metrics

The performance of our method was computed and tested using the following figures of merit. As mentioned above, the performance has been evaluated by five-fold cross validation as follows:

1) Sensitivity (or recall) is the coverage of positives i.e. the percent of correctly predicted Ca^{2+}-binding 12-mers and correct estimation of their affinity.

$$Sensitivity = [TP/(TP+FN)] \times 100$$

2) Specificity is the coverage of negatives, that is, the percent of correctly predicted Ca^{2+} non-binding 12-mers and correct estimation of their affinity.

$$Specificity = [TN/(TN+FP)] \times 100$$

3) Accuracy is the percentage of correctly predicted positives and negatives.

$$Accuracy = [(TP+TN)/(TP+FP+TN+FN)] \times 100$$

4) MCC – Matthews's correlation coefficient is the statistical parameter to assess the quality of the prediction and account for unbalancing in data [58]. An MCC equal to 1 is regarded as a perfect prediction, whereas that equal to 0 indicates a completely random prediction.

$$MCC = (TP)(TN) - (FP)(FN)/\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}$$

[TP = true positive; FN = false negative; TN = true negative; FP = false positive]

5) AUC (Area under the ROC Curve) – Receiver Operating Curve (ROC) and AUC were computed using SPSS software. It generates ROC curves and calculates AUC by ranking the decision values.

Supporting Information

File S1 File S1 includes the following: Figure S1. a) Plot of affinity vs. PSSM for the test data set (D5). The calculated correlation coefficient obtained was 0.61 using [41] amino acid

frequencies. **Figure S2.** The isothermal titration calorimetric analysis of Ca^{2+}-binding to apo-EhCaBPs.ITC experiments were carried out as described under "Materials and Methods". Plot of heat absorbed/released (In kcal mol^{-1}) per injection of $CaCl_2$ as a function of molar ratio of Ca^{2+}: protein at 25°C is shown. For all titrations, the top panels represent the raw data (power: time) and the bottom panels represent integrated binding isotherms. The solid line represents the best nonlinear fit to the experimental data. Binding isotherm for A: EhCaBP3; B: EhCaBP4; C: EhCaBP5; D: EhCaBP6 and E: EhCaBP7. Thermodynamic parameters obtained are summarized in Table 1. **Figure S3.** ROC plots of AC&CC, AC&HC, AC&HC&HYC, AC&HYC&CC and AC&HYC for the datasets D5–D7 set. Receiver operating characteristic (ROC) plot used for depicting relative trade-offs between true positive and false positives. The corresponding AUC value of each model is shown in brackets. **Figure S4.** Schematic representation of the procedure for model development and feature selection for EF-hand loop region prediction and estimation of binding affinity and its web implementation. The procedure is explained in detail in the "Methods" section. A). A group of sequences with known EF-hand structural motifs were downloaded and further classified into two groups after removing redundant sequences using CD-HIT. The sequences were further converted into binary and amino acid composition (AAC) profiles for SVM input. Models were generated using LIBSVM and were tested on all the datasets (D3–D6) and further validated by scanning the E. *histolytica* proteome. B). Non-redundant sequences of EF-hand loops from known structures were classified into two groups on the basis of scores obtained from position-specific scoring metrics. The sequences were then converted into binary, AAC and different amino acid indices patterns. We have generated both standalone and combinations of features (2, 3, 4, 5) using a Perl script written in-house. The input vectors were trained using LIBSVM and cudized LIBSVM and selected on the basis of their performance on experimental datasets using 5-fold cross validation accuracy threshold >70 %. The best performing models selected from screening were further validated using three different experimentally derived datasets on EF hand motifs. The final step involved web implementation of the best (AC&HC) model. **Table S1.** The χ^2 value for each amino acid residue is estimated with one degree of freedom and significance level $P = 0.001$. The $\Sigma\chi^2$ values are estimated with 19 degrees of freedom and significance level $P<0.001$. The expected (Exp) and observed (Obs) values and the corresponding χ^2 values for amino acid residues and the $\Sigma\chi^2$ values for those positions that do not reach 10.8 and 43.8 (for one and 19 degrees of freedom, respectively) are given more significance. **Table S2.** Test Dataset: Summary of EF hand loops obtained from the literature and their macroscopic binding constant along with CAL-EF-AFi predictions (D5). The classification details with supportive binding constants are listed under "Author's Note". (Red-colored affinities are the false negative affinity predictions, and turquoise-colored sequences are the false negative EF loop predictions). **Table S3.** Independent dataset (D6) summary of EF hand loops obtained from Boguta, et al., 1988 [59]. The table contains average binding constants of Ca^{2+} for troponin C superfamily (TnC) proteins from experimental data reported by various laboratories. The classification details with supportive binding constants are listed under "Author's Note". (Red-colored affinities are the false positive predictions). **Table S4.** Validation dataset summary of EF-hand loops obtained from ITC studies of CaBPs from *E. histolytica* and their macroscopic binding constant according to CAL-EF-AFi's predictions (D7). The classification details with supportive binding constants are listed under "Author's Note" (Red-colored affinities are the false positive predictions). **Table S5.** Predictions of putative EF hand-containing calcium-binding protein and their calcium-binding affinities from the *E. histolytica* proteome. **Table S6.** The performance and comparison of CAL-EF-AFi with PFAM and Calpred on the E. *histolytica* proteome. Listed are the sequences predicted by CAL-EF-AFi followed by PFAM-based HMM model prediction and CalPred's predictions. (Legends for CAL-EF-AFi's prediction: number of Ca^{2+}-binding loop sequence prediction, residue number followed by sequence and SVM scores; Legends for PFAM predictions: red-colored region is the loop region predicted, followed by the E-value for the sequence; Legends for CalPred predictions: X: Non-Binding region C: Calcium Binding region). **Table S7.** Calcium-binding EF-hand protein sequences in FASTA format at 60% sequence redundancy with EF-hand loop region residues labeled in lower case letters. **(D1). Table S8.** The list of 12-mer sequences from non-binding regions of calcium-binding EF-hand proteins greater than 60% sequence redundancy. **Table S9.** The training data used for estimation of binding affinity were taken from the RCSB based on PSSM scores obtained from the EF-hand loop region. The positive dataset (**D3**) consisted of one hundred forty four 12-mer sequences and there were 124 sequences in the negative dataset (**D4**). **Table S10.** The redundant set of PDB ids of EF hand-containing calcium-binding proteins. The sequences taken from the RCSB were further processed using CD-HIT and the list if the sequences with different threshold are listed in Table S11. **Table S11.** The sequence-wise classification of data obtained from PROSITE and RCSB- The data was further processed by using CD-HIT at 90%, 70%, 60%, 50% sequence redundancy cutofffor classification of EF-hand loop Ca^{2+}-binding and non-binding region.
(DOC)

Acknowledgments

The authors thank Dr. Jerry Brown, Brandeis University for critically proofreading and editing the manuscript and Mr Vineet Jha for helping with the coding.

Author Contributions

Conceived and designed the experiments: MM AB SG. Performed the experiments: MM NP. Analyzed the data: MM NP AB SG. Contributed reagents/materials/analysis tools: AB SG. Wrote the paper: MM NP AB SG.

References

1. Berridge MJ, Bootman MD, Lipp P (1998) Calcium – a life and death signal. Nature 395: 645–648.
2. Ermak G, Davies KJ (2002) Calcium and oxidative stress: from cell signaling to cell death. Mol Immunol 38: 713–721.
3. Verkhratsky A (2007) Calcium and cell death. Subcell Biochem 45: 465–480.
4. Bencina M, Bagar T, Lah L, Krasevec N (2009) A comparative genomic analysis of calcium and proton signaling/homeostasis in Aspergillus species. Fungal Genet Biol 46 Suppl 1: S93–S104.
5. Gangola P, Rosen BP (1987) Maintenance of intracellular calcium in Escherichia coli. J Biol Chem 262: 12570–12574.
6. Zhou Y, Frey TK, Yang JJ (2009) Viral calciomics: interplays between Ca2+ and virus. Cell Calcium 46: 1–17.
7. Herzberg O, Moult J, James MN (1986) A model for the Ca2+-induced conformational transition of troponin C. A trigger for muscle contraction. J Biol Chem 261: 2638–2644.
8. Holmes KC, Popp D, Gebhard W, Kabsch W (1990) Atomic model of the actin filament. Nature 347: 44–49.
9. Mann KG, Nesheim ME, Church WR, Haley P, Krishnaswamy S (1990) Surface-dependent reactions of the vitamin K-dependent enzyme complexes. Blood 76: 1–16.

10. Carafoli E (2002) Calcium signaling: a tale for all seasons. Proc Natl Acad Sci U S A 99: 1115–1122.
11. Sutton RB, Davletov BA, Berghuis AM, Sudhof TC, Sprang SR (1995) Structure of the first C2 domain of synaptotagmin I: a novel Ca2+/phospholipid-binding fold. Cell 80: 929–938.
12. Henikoff S, Greene EA, Pietrokovski S, Bork P, Attwood TK, et al. (1997) Gene families: the taxonomy of protein paralogs and chimeras. Science 278: 609–614.
13. Kawasaki H, Nakayama S, Kretsinger RH (1998) Classification and evolution of EF-hand proteins. Biometals 11: 277–295.
14. Grabarek Z (2006) Structural basis for diversity of the EF-hand calcium-binding proteins. J Mol Biol 359: 509–525.
15. Bairoch A, Cox JA (1990) EF-hand motifs in inositol phospholipid-specific phospholipase C. FEBS Lett 269: 454–456.
16. Finn BE, Forsen S (1995) The evolving model of calmodulin structure, function and activation. Structure 3: 7–11.
17. Nakayama S, Kretsinger RH (1994) Evolution of the EF-hand family of proteins. Annu Rev Biophys Biomol Struct 23: 473–507.
18. Gifford JL, Walsh MP, Vogel HJ (2007) Structures and metal-ion-binding properties of the Ca2+-binding helix-loop-helix EF-hand motifs. Biochem J 405: 199–221.
19. Godzik A, Sander C (1989) Conservation of residue interactions in a family of Ca-binding proteins. Protein Eng 2: 589–596.
20. Linse S, Brodin P, Johansson C, Thulin E, Grundstrom T, et al. (1988) The role of protein surface charges in ion binding. Nature 335: 651–652.
21. Linse S, Forsen S (1995) Determinants that govern high-affinity calcium binding. Adv Second Messenger Phosphoprotein Res 30: 89–151.
22. Lin HH, Han LY, Zhang HL, Zheng CJ, Xie B, et al. (2006) Prediction of the functional class of metal-binding proteins from sequence derived physicochemical properties by support vector machine approach. BMC Bioinformatics 7 Suppl 5: S13.
23. Zhou Y, Yang W, Kirberger M, Lee HW, Ayalasomayajula G, et al. (2006) Prediction of EF-hand calcium-binding proteins and analysis of bacterial EF-hand proteins. Proteins 65: 643–655.
24. Franke S, Herfurth J, Hoffmann D (2010) Estimating affinities of calcium ions to proteins. Adv Appl Bioinform Chem 3: 1–6.
25. Boguta G, Stepkowski D, Bierzynski A (1988) Theoretical estimation of the calcium-binding constants for proteins from the troponin C superfamily based on a secondary structure prediction method. II. Applications. J Theor Biol 135: 63–73.
26. Wiseman T, Williston S, Brandts JF, Lin LN (1989) Rapid measurement of binding constants and heats of binding using a new titration calorimeter. Anal Biochem 179: 131–137.
27. Bhattacharya A, Padhan N, Jain R, Bhattacharya S (2006) Calcium-binding proteins of Entamoeba histolytica. Arch Med Res 37: 221–225.
28. Kunal J, Chandan K, PK N (2010) Prediction of EF-hand calcium-binding proteins and identification of calcium-binding regions using machine learning techniques. Journal of Cell and Molecular Biology 8(2): 41–49.
29. VanScyoc WS, Sorensen BR, Rusinova E, Laws WR, Ross JB, et al. (2002) Calcium binding to calmodulin mutants monitored by domain-specific intrinsic phenylalanine and tyrosine fluorescence. Biophys J 83: 2767–2780.
30. Moeschler HJ, Schaer JJ, Cox JA (1980) A thermodynamic analysis of the binding of calcium and magnesium ions to parvalbumin. Eur J Biochem 111: 73–78.
31. Deng H, Chen G, Yang W, Yang JJ (2006) Predicting calcium-binding sites in proteins – a graph theory and geometry approach. Proteins 64: 34–42.
32. Wang X, Kirberger M, Qiu F, Chen G, Yang JJ (2009) Towards predicting Ca2+-binding sites with different coordination numbers in proteins with atomic resolution. Proteins 75: 787–798.
33. Liu T, Altman RB (2009) Prediction of calcium-binding sites by combining loop-modeling with machine learning. BMC Struct Biol 9: 72.
34. Bock JR, Gough DA (2001) Predicting protein – protein interactions from primary structure. Bioinformatics 17: 455–460.
35. Lo SL, Cai CZ, Chen YZ, Chung MC (2005) Effect of training datasets on support vector machine prediction of protein-protein interactions. Proteomics 5: 876–884.
36. Rout AK, Padhan N, Barnwal RP, Bhattacharya A, Chary KV (2010) Calmodulin-like Protein from Entamoeba histolytica: Solution Structure and Calcium-Binding Properties of a Partially Folded Protein. Biochemistry.
37. Gopal B, Swaminathan CP, Bhattacharya S, Bhattacharya A, Murthy MR, et al. (1997) Thermodynamics of metal ion binding and denaturation of a calcium binding protein from Entamoeba histolytica. Biochemistry 36: 10910–10916.
38. Sigrist CJ, de Castro E, Cerutti L, Cuche BA, Hulo N, et al. (2013) New and continuing developments at PROSITE. Nucleic Acids Res 41: D344–347.
39. Berman H, Henrick K, Nakamura H, Markley JL (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res 35: D301–303.
40. Li W, Jaroszewski L, Godzik A (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics 17: 282–283.
41. Marsden BJ, Shaw GS, Sykes BD (1990) Calcium binding proteins. Elucidating the contributions to calcium affinity from an analysis of species variants and peptide fragments. Biochemistry and Cell Biology 68: 587–601.
42. Gribskov M, McLachlan AD, Eisenberg D (1987) Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci U S A 84: 4355–4358.
43. Henikoff JG, Henikoff S (1996) Using substitution probabilities to improve position-specific scoring matrices. Comput Appl Biosci 12: 135–143.
44. Tatusov RL, Altschul SF, Koonin EV (1994) Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. Proc Natl Acad Sci U S A 91: 12091–12095.
45. Eddy SR (2004) Where did the BLOSUM62 alignment score matrix come from? Nat Biotechnol 22: 1035–1036.
46. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, et al. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci U S A 97: 262–267.
47. Ding CH, Dubchak I (2001) Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 17: 349–358.
48. Chang C-CaL, Chih-Jen (2011) LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2: 27:21–27:27.
49. A Athanasopoulos AD, Mezaris V, Kompatsiaris I (April 2011) GPU Acceleration for Support Vector Machines. Proc 12th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2011).
50. Xiao X, Shao S, Ding Y, Huang Z, Chou KC (2006) Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. Amino Acids 30: 49–54.
51. Xiao X, Wang P, Chou KC (2009) GPCR-CA: A cellular automaton image approach for predicting G-protein-coupled receptor functional classes. J Comput Chem 30: 1414–1423.
52. Bhasin M, Raghava GP (2004) Analysis and prediction of affinity of TAP binding peptides using cascade SVM. Protein Sci 13: 596–607.
53. Ramana J, Gupta D (2010) FaaPred: a SVM-based prediction method for fungal adhesins and adhesin-like proteins. PLoS One 5: e9695.
54. Klein P, Kanehisa M, DeLisi C (1984) Prediction of protein function from sequence properties. Discriminant analysis of a data base. Biochim Biophys Acta 787: 221–226.
55. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157: 105–132.
56. Kuhn LA, Swanson CA, Pique ME, Tainer JA, Getzoff ED (1995) Atomic and residue hydrophilicity in the context of folded protein structures. Proteins 23: 536–547.
57. Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH (1985) Hydrophobicity of amino acid residues in globular proteins. Science 229: 834–838.
58. Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405: 442–451.
59. Boguta G, Stepkowski D, Bierzynski A (1988) Theoretical estimation of the calcium-binding constants for proteins from the troponin C superfamily based on a secondary structure prediction method. I. Estimation procedure. J Theor Biol 135: 41–61.

Hindawi
PPAR Research
Volume 2017, Article ID 6397836, 11 pages
https://doi.org/10.1155/2017/6397836

Research Article

Investigations on Binding Pattern of Kinase Inhibitors with PPARγ: Molecular Docking, Molecular Dynamic Simulations, and Free Energy Calculation Studies

Mohit Mazumder,[1] Prija Ponnan,[2] Umashankar Das,[2] Samudrala Gourinath,[1] Haseeb Ahmad Khan,[3] Jian Yang,[2] and Meena Kishore Sakharkar[2]

[1] *Structural Biology Laboratory, School of Life Sciences, Jawaharlal Nehru University, New Delhi, India*
[2] *Drug Discovery and Development Research Group, College of Pharmacy and Nutrition, University of Saskatchewan, 107 Wiggins Road, Saskatoon, SK, Canada S7N 5C9*
[3] *Department of Biochemistry, College of Science, King Saud University, Riyadh, Saudi Arabia*

Correspondence should be addressed to Meena Kishore Sakharkar; meena.sakharkar@usask.ca

Received 2 November 2016; Accepted 4 January 2017; Published 22 February 2017

Academic Editor: Constantinos Giaginis

Peroxisome proliferator-activated receptor gamma (PPARγ) is a potential target for the treatment of several disorders. In view of several FDA approved kinase inhibitors, in the current study, we have investigated the interaction of selected kinase inhibitors with PPARγ using computational modeling, docking, and molecular dynamics simulations (MDS). The docked conformations and MDS studies suggest that the selected KIs interact with PPARγ in the ligand binding domain (LBD) with high positive predictive values. Hence, we have for the first time shown the plausible binding of KIs in the PPARγ ligand binding site. The results obtained from these in silico investigations warrant further evaluation of kinase inhibitors as PPARγ ligands in vitro and in vivo.

1. Introduction

Peroxisome proliferator-activated receptors (PPARs) belong to the nuclear receptor super family and are ligand activated transcription factors, regulating the expression of a wide variety of genes. On activation by a ligand, they bind to the PPAR-responsive regulatory elements (PPRE) and/or PPAR associated conserved motif (PACM) as obligate heterodimers with retinoid X receptor (RXR) [1, 2]. Similar to other nuclear receptor-family members, PPARs are multidomain proteins, consisting of an N-terminal transactivation domain (AF1), a highly conserved DNA-binding domain (DBD), and a C-terminal ligand binding domain (LBD) which has a ligand-dependent transactivation function (AF2) [3, 4]. Three isoforms of PPARs (alpha, beta/delta, and gamma) have been identified so far in human, mouse, rats, xenopus, and hamsters [5–7] and among them, PPARγ is the most intensively studied. PPARγ has three alternatively spliced isoforms and all of them are expressed in adipose tissues [8, 9]. It is primarily involved in the regulation of lipid metabolism and insulin sensitivity reactions and also plays an important role in carcinogenesis and cell physiology [10, 11]. Also, PPARs have been shown to have ligand independent repression whereby they repress the transcription of direct target genes by recruitment of corepressor complexes which blocks the actions of coactivator complexes [12]. PPARγ activation is involved in transcriptional regulation of genes involved in proliferation, angiogenesis, apoptosis, organogenesis, and energy metabolism and hence implicated in cell growth and viability [13–16]. PPARγ signaling is modulated using different domains and various natural lipophilic agonists (ligands) such as unsaturated fatty acids, oxidized lipid species, eicosanoids, and prostaglandins [2, 17, 18]. Conformational changes caused by ligand binding lead to the modulation of PPARγ activity by differential recruitment of cofactors [4, 12]. PPARγ exhibits high affinity towards thiazolidinediones (TZDs) [19]. TZDs including troglitazone,

SCIENTIFIC REPORTS

OPEN

Received: 29 June 2016
Accepted: 14 October 2016
Published: 03 November 2016

N-acetyl ornithine deacetylase is a moonlighting protein and is involved in the adaptation of *Entamoeba histolytica* to nitrosative stress

Preeti Shahi[1], Meirav Trebicz-Geffen[1], Shruti Nagaraja[1], Rivka Hertz[1], Sharon Baumel-Alterzon[1], Karen Methling[2], Michael Lalk[2], Mohit Mazumder[3], Gourinath Samudrala[3] & Serge Ankri[1]

Adaptation of the *Entamoeba histolytica* parasite to toxic levels of nitric oxide (NO) that are produced by phagocytes may be essential for the establishment of chronic amebiasis and the parasite's survival in its host. In order to obtain insight into the mechanism of *E. histolytica*'s adaptation to NO, *E. histolytica* trophozoites were progressively adapted to increasing concentrations of the NO donor drug, S-nitrosoglutathione (GSNO) up to a concentration of 110 μM. The transcriptome of NO adapted trophozoites (NAT) was investigated by RNA sequencing (RNA-seq). N-acetyl ornithine deacetylase (NAOD) was among the 208 genes that were upregulated in NAT. NAOD catalyzes the deacetylation of N-acetyl-L-ornithine to yield ornithine and acetate. Here, we report that NAOD contributes to the better adaptation of the parasite to nitrosative stress (NS) and that this function does not depend on NAOD catalytic activity. We also demonstrated that glyceraldehyde 3-phosphate dehydrogenase (GAPDH) is detrimental to *E. histolytica* exposed to NS and that this detrimental effect is neutralized by NAOD or by a catalytically inactive NAOD (mNAOD). These results establish NAOD as a moonlighting protein, and highlight the unexpected role of this metabolic enzyme in the adaptation of the parasite to NS.

Intestinal infections are a global medical problem and diarrheal disease is one of the main causes of childhood morbidity and mortality. *Entamoeba histolytica* is a protozoan parasite and the causal agent of amebiasis, the second most common cause of death from parasitic disease worldwide after malaria (at least 100,000 deaths each year). According to the World Health Organization, amebic dysentery affects 50 million people in India, Southeast Asia, Africa, and Latin America. Since poor sanitary conditions and unsafe hygiene practices exist in many parts of the world, the main mode of transmission of amebiasis is the ingestion of food and/or water that is contaminated with feces and *E. histolytica* cysts. *E. histolytica* trophozoites are non-pathogenic commensals in 90% of infected individuals (asymptomatic amebiasis). For unknown reasons, some of these trophozoites can invade the intestinal mucosa, cause dysentery, and migrate to the liver where they produce abscesses (extraintestinal amebiasis). In the large intestine, *E. histolytica* is exposed to nanomolar concentrations of nitric oxide (NO) that is produced in intestinal epithelial cells by constitutive NO synthase (NOS)[1] and as an intermediate in denitrification by the intestinal microbiota[2]. Although exposure to low NO concentrations is insufficient to kill the parasite[3], these low concentrations may strengthen its resistance to high NO concentrations. Amebiasis is characterized by acute inflammation of the intestine with the release of cytokines, such as tumor necrosis factor α (TNFα), interleukin 8 (IL-8), interferon gamma (IFN-γ), and interleukin β (IL-1β), and the generation of micromolar concentrations of reactive oxygen species (ROS) and reactive nitrogen species (RNS) from activated cells of the host's immune system (for a recent review see ref. 4). NO in micromolar concentrations is

[1]Department of Molecular Microbiology, Ruth and Bruce Rappaport Faculty of Medicine, Technion, P.O.B. 9649, 31096 Haifa Israel. [2]University of Greifswald, Institute of Biochemistry, Greifswald, Germany. [3]Jawaharlal Nehru University School of Life Sciences, New Delhi, India. Correspondence and requests for materials should be addressed to M.T.-G. (email: meiravg@tx.technion.ac.il) or S.A. (email: sankri@tx.technion.ac.il)

SCIENTIFIC REPORTS

OPEN

Structural insight into β-Clamp and its interaction with DNA Ligase in *Helicobacter pylori*

Received: 28 April 2016
Accepted: 14 July 2016
Published: 08 August 2016

Preeti Pandey[1,2], Khaja Faisal Tarique[1], Mohit Mazumder[1], Syed Arif Abdul Rehman[1], Nilima kumari[2] & Samudrala Gourinath[1]

***Helicobacter pylori*, a gram-negative and microaerophilic bacterium, is the major cause of chronic gastritis, gastric ulcers and gastric cancer. Owing to its central role, DNA replication machinery has emerged as a prime target for the development of antimicrobial drugs. Here, we report 2Å structure of β-clamp from *H. pylori* (Hpβ-clamp), which is one of the critical components of DNA polymerase III. Despite of similarity in the overall fold of eubacterial β-clamp structures, some distinct features in DNA interacting loops exists that have not been reported previously. The *in silico* prediction identified the potential binders of β-clamp such as alpha subunit of DNA pol III and DNA ligase with identification of β-clamp binding regions in them and validated by SPR studies. Hpβ-clamp interacts with DNA ligase in micromolar binding affinity. Moreover, we have successfully determined the co-crystal structure of β-clamp with peptide from DNA ligase (not reported earlier in prokaryotes) revealing the region from ligase that interacts with β-clamp.**

The sliding clamp is a ring-shaped protein complex that encircles DNA with the help of clamp loader in an ATP-dependent manner, and slides along the DNA. Because of its ability to slide along DNA, the sliding clamp is required by many different enzymes for DNA replication and repair[1]. Clamps not only increase the processivity of these enzymes but also serve as attachment points to coordinate their activities. The clamps are thus required for keeping these enzymes tightly associated with DNA while at the same time facilitating their translocation along duplex DNA[2].

The elongation factor β-clamp also called sliding clamp has been found to exist in both prokaryotes and eukaryotes. In eukaryotes, it is generally known by the name PCNA and is a heterotrimer. Each monomer consists of two domains, with N-terminal domain joint to C-terminal domain of neighboring monomers by non-covalent interactions and form a ring-shaped structure[3]. In prokaryotes, however, β-clamp is a homodimer, with each monomer consisting of three globular domains, and in this way β-clamp displays a six-domain ring[4]. Thus despite of having sequence similarity between these two, they share similar architecture as suggested by their structural analysis[3,5]. All of the known clamp-binding proteins contain a conserved peptide sequence motif through which they interact with the clamp[6]. In both prokaryotes and eukaryotes, a key feature of this clamp-binding motif is the presence of hydrophobic amino acid residues that bind to the hydrophobic pocket in the C-terminal region of the clamp. Based on experimental studies, QL(S/D)LF[7] and QxxL(x)F[8] have been proposed as consensus binding sequences for *E. coli*β-clamp.

Although β-clamp is part of the DNA polymerase III holoenzyme, it is not attached to polymerase III permanently like the other subunits. β-clamp is loaded on the DNA, by clamp loader, a subunit of DNA Pol III. It interacts with several proteins other than DNA polymerase III subunits; it also freely slides along DNA and improves the processivity of other proteins. Among the several β-clamp-interacting partners, one of the most important protein is DNA ligase. After completion of the synthesis of the lagging strand fragment, DNA polymerase III becomes separated from β-clamp and DNA and moves to another primed site[9]. It was hypothesized that this released beta clamp interacts with Pol I, which digests RNA primer at the 5′ end of the primer and replaces it with DNA by nick translation. After that, the clamp interacts with DNA ligase, which seals the nick[10]. Thus, the interaction between β-clamp and DNA ligase helps in Okazaki fragment maturation, and is also needed for DNA repair. Therefore, studying the interaction between these two components is of great importance. In case of

[1]School of Life Sciences, Jawaharlal Nehru University, New Delhi, India. [2]Department of Bioscience and Biotechnology, Banasthali University, Rajasthan, India. Correspondence and requests for materials should be addressed to S.G. (email: sgourinath@mail.jnu.ac.in)

FEBS Letters

FEBS PRESS
science publishing by scientists

Crystal structure of *Arabidopsis thaliana* calmodulin7 and insight into its mode of DNA binding

Sanjeev Kumar[1], Mohit Mazumder[1], Nisha Gupta[2], Sudip Chattopadhyay[2] and Samudrala Gourinath[1]

1 School of Life Sciences, Jawaharlal Nehru University, New Delhi, India
2 Department of Biotechnology, National Institute of Technology, Durgapur, India

Correspondence
S. Gourinath, School of Life Sciences, Jawaharlal Nehru University, New Delhi 110067, India
Fax: +1 91-11-26742916/2558
Tel: +1 91-11-26704513
E-mail: sgourinath@mail.jnu.ac.in

(Received 14 May 2016, revised 6 July 2016, accepted 1 August 2016, available online 24 August 2016)

doi:10.1002/1873-3468.12349

Edited by Richard Cogdell

Calmodulin (CaM) is a Ca^{2+} sensor that participates in several cellular signaling cascades by interacting with various targets, including DNA. It has been shown that *Arabidopsis thaliana* CaM7 (AtCaM7) interacts with Z-box DNA and functions as a transcription factor [Kushwaha R *et al.* (2008) *Plant Cell* 20, 1747–1759; Abbas N *et al.* (2014) *Plant Cell* 26, 1036–1052]. The crystal structure of AtCaM7, and a model of the AtCAM7-Z-box complex suggest that Arg-127 determines the DNA-binding ability by forming crucial interactions with the guanine base. We validated the model using biolayer interferometry, which confirmed that AtCaM7 interacts with Z-box DNA with high affinity. In contrast, the AtCaM2/3/5 isoform does not show any binding, although it differs from AtCaM7 by only a single residue.

Keywords: CaM; molecular modeling; protein crystallization; protein–DNA interaction

Calmodulin (CaM) are ubiquitous eukaryotic proteins that can bind to a variety of protein targets in response to Ca^{2+} signals. CaM plays essential role in Ca^{2+} signaling, regulating numerous intracellular processes such as cell motility, growth, proliferation, and apoptosis [1]. CaM binds Ca^{2+} ions using its helix-loop-helix (EF-hand) structural motif, and this motif generally undergoes large conformational changes upon Ca^{2+} binding [2]. In the Ca^{2+}-loaded form, the CaM adopts stable state, and each EF-hand opens so that its two alpha helices become perpendicular to each other. In the Ca^{2+}-free (apo) form, CaM adopts a closed and flexible state where EF-hand motifs are in closed conformation [3–5]. In contrast to apo form, Ca^{2+}-loaded CaM binds to many (> 300) target proteins that regulates the various biological processes [6–8].

CaM responds to a wide range of Ca^{2+} concentrations (10^{-12} to 10^{-6} M) in Ca^{2+}-dependent signal transduction, after binding of Ca^{2+} to EF hand motifs, CaM can bind to different target proteins to accomplish these physiological roles [9,10]. CaMs are found to be involved in various signaling event and these signaling are governed in Ca^{+2}-dependent as well as Ca^{2+}-independent manners. In many cases these signaling mechanisms have been elucidated on structural basis in from of protein–protein (CaM complex with its target proteins) complex structures [11,12]. CaM can regulate basic helix-loop-helix transcription factors where CaM inhibits DNA–protein interactions by competing with the DNA-binding domains of the basic helix-loop-helix proteins [13]. Helix-loop-helix motifs, such as that in the EF-hand-containing protein DREM, have been reported to interact with DNA [14–16]. Aside from interactions with various proteins, CaM can also interact with DNA and serve as transcription factors.

The *A. thaliana* genome contains seven CaM genes that encode four protein isoforms: CaM1/CaM4,

Abbreviations
BLI, biolayer interferometry; CaM, calmodulin.

Structure-Based Design of Inhibitors of the Crucial Cysteine Biosynthetic Pathway Enzyme O-Acetyl Serine Sulfhydrylase

Mohit Mazumder and Samudrala Gourinath*

School of life sciences, Jawaharlal Nehru University, N.Delhi-110067, India

Please provide corresponding author(s) photograph
size should be 4" x 4" inches

Abstract: The cysteine biosynthetic pathway is of fundamental importance for the growth, survival, and pathogenicity of the many pathogens. This pathway is present in many species but is absent in mammals. The ability of pathogens to counteract the oxidative defences of a host is critical for the survival of these pathogens during their long latent phases, especially in anaerobic pathogens such as *Entamoeba histolytica, Leishmania donovani, Trichomonas vaginalis,* and *Salmonella typhimurium*. All of these organisms rely on the *de novo* cysteine biosynthetic pathway to assimilate sulphur and maintain a ready supply of cysteine. The *de novo* cysteine biosynthetic pathway, on account of its being important for the survival of pathogens and at the same time being absent in mammals, is an important drug target for diseases such as amoebiasis, trichomoniasis & tuberculosis. Cysteine biosynthesis is catalysed by two enzymes: serine acetyl transferase (SAT) followed by O-acetylserine sulfhydrylase (OASS). OASS is well studied, and with the availability of crystal structures of this enzyme in different conformations, it is a suitable template for structure-based inhibitor development. Moreover, OASS is highly conserved, both structurally and sequence-wise, among the above-mentioned organisms. There have been several reports of inhibitor screening and development against this enzyme from different organisms such as *Salmonella typhimurium, Mycobacterium tuberculosis* and *Entamoeba histolytica*. All of these inhibitors have been reported to display micromolar to nanomolar binding affinities for the open conformation of the enzyme. In this review, we highlight the structural similarities of this enzyme in different organisms and the attempts for inhibitor development so far. We also propose that the intermediate state of the enzyme may be the ideal target for the design of effective high-affinity inhibitors.

Keywords: Conformational changes, cysteine biosynthetic pathway, Inhibitors, O-acetyl serine sulfhydrylase, pathogens.

1. INTRODUCTION

Cysteine plays a vital role in organic sulphur metabolism. Utilisation of sulphur from cysteine is the initial step of many biosynthetic pathways that supply the cell with biomolecules such as Fe–S clusters, modified tRNAs (thiouridine), thiamine, biotin, glutathione, trypanothione and mycothiol (Beinert, 2000;Kessler, 2006). Mammals rely on sulphated amino acids, mostly the essential amino acid methionine, for their sulphur supplies whereas most bacteria, protists and plants assimilate sulphur into cysteine through the reductive sulphate assimilation pathway (RSAP). Apart from functioning as a protein building block and as a component of important biomolecules, cysteine participates directly, or as a precursor of reducing agents, in the maintenance of the redox state of the cell. This function is of special interest to microorganisms that spend part of their life cycle in highly oxidizing environments, e.g., when establishing an infection in the human host or inside human macrophages (Mozzarelli *et al.*, 2011).

The ability of pathogens to counteract the oxidative defences of a host is critical for the survival of these pathogens during their long latent phases, especially in anaerobic pathogens such as *Entamoeba histolytica* and pathogens such as *Leishmania donovani, Trichomonas vaginalis,* and *Salmonella typhimurium* poses a severe threat to health. Due to the increase in drug resistance, treatment of such diseases have become complicated and hence life threating. All of these organisms rely on the *de novo* cysteine biosynthetic pathway to assimilate sulphur and maintain a ready supply of cysteine (Campanini *et al.*, 2015). The *de novo* cysteine biosynthetic pathway, on account of its being important for the pathogen and at the same time being absent in mammals, is an important drug target. An infection by these pathogens cause diseases such as amebiasis, leishmaniosis and tuberculosis where the inhibitors of cysteine biosynthetic pathway could result in better treatment then the presently available antibiotics.

In cysteine biosynthesis, the first reaction is catalysed by the enzyme serine acetyltransferase (SAT, EC 2.3.1.30), which generates the activated sulphide acceptor O-acetylserine (OAS) from serine and acetyl CoA. In the second step, O-acetylserine sulfhydrylase (OASS, EC 2.5.1.47) catalyzes O-acetylserine where sulphide is inserted into O-acetylserine in a β replacement reaction catalyzed by the cofactor pyridoxal phosphate (PLP) to yield cysteine and acetate (Fig. **1**). The enzymatic pathway of cysteine synthesis was characterized by the pioneering work of

*Address correspondence to this author at the School of Life Sciences, Jawaharlal Nehru University, New Delhi 110 067, India;
Tel: ????????????????; Fax: ????????????????;
E-mail: samudralag@yahoo.com

1568-0266/16 $58.00+.00

Biochimica et Biophysica Acta 1860 (2016) 1508–1518

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbagen

Structural investigation and inhibitory response of halide on phosphoserine aminotransferase from *Trichomonas vaginalis*

Rohit Kumar Singh, Mohit Mazumder, Bhumika Sharma, Samudrala Gourinath *

School of Life Sciences, Jawaharlal Nehru University, New Delhi, India

ARTICLE INFO

Article history:
Received 12 December 2015
Received in revised form 4 April 2016
Accepted 17 April 2016
Available online 19 April 2016

Keywords:
Serine pathway
Phosphoserine aminotransferase
Structure
Enzyme kinetics
Inhibition by halides
Molecular dynamics simulation

ABSTRACT

Background: Phosphoserine aminotransferase (PSAT) catalyses the second reversible step of the phosphoserine biosynthetic pathway in *Trichomonas vaginalis*, which is crucial for the synthesis of serine and cysteine.
Methods: PSAT from *T. vaginalis* (TvPSAT) was analysed using X-ray crystallography, enzyme kinetics, and molecular dynamics simulations.
Results: The crystal structure of TvPSAT was determined to 2.15 Å resolution, and is the first protozoan PSAT structure to be reported. The active site of TvPSAT structure was found to be in a closed conformation, and at the active site PLP formed an internal aldimine linkage to Lys 202. In TvPSAT, Val 340 near the active site while it is Arg in most other members of the PSAT family, might be responsible in closing the active site. Kinetic studies yielded Km values of 54 μM and 202 μM for TvPSAT with OPLS and AKG, respectively. Only iodine inhibited the TvPSAT activity while smaller halides could not inhibit.
Conclusion: Results from the structure, comparative molecular dynamics simulations, and the inhibition studies suggest that iodine is the only halide that can bind TvPSAT strongly and may thus inhibit the activity of TvPSAT. The long loop between β8 and α8 at the opening of the TvPSAT active site cleft compared to other PSATs, suggests that this loop may help control the access of substrates to the TvPSAT active site and thus influences the enzyme kinetics.
General significance: Our structural and functional studies have improved our understanding of how PSAT helps this organism persists in the environment.

1. Introduction

Trichomonas vaginalis is the causative agent of human trichomoniasis [1], which is the most common non-viral sexually transmitted disease in the world. *T. vaginalis* causes infection in both men and women, but less frequently in men. Symptoms and consequences of this infection are generally mild, and include problems during pregnancy as well as premature birth and low birth weight [2,3]. *T. vaginalis* is adapted to environments containing low concentrations of oxygen by being a fundamentally fermentative organism [4,5], and has to withstand any oxidizing conditions encountered in order to survive. Various metabolites likely to arise from the metabolism of oxygen (such as H_2O_2,OH free radical and the superoxide radical anion) are generally harmful to cells and so need to be countered, but trichomonads such as *T. vaginalis* lack glutathione (an antioxidant found in most eukaryotes) and related thiols [6]. However, cysteine has been generally believed to be the major cellular reducing agent and antioxidant [7] in most protozoans.

In *T. vaginalis*, production of cysteine is heavily dependent on the phosphorylated serine metabolic pathway. This pathway is associated with the production of L-serine in many organisms including bacteria, yeast, plants and animals [8–11], and consists of the sequential reactions catalysed by D-phosphoglycerate dehydrogenase (PGDH, EC 1.1.1.95), phosphoserine aminotransferase (PSAT, EC 2.6.1.52), and O-phosphoserine phosphatase (PSP, EC 3.1.3.3) (Fig. 1). In the first committed step of this pathway, PGDH catalyses the synthesis of phosphohydroxypyruvate (PHP) from the glycolytic intermediate D-3-phosphoglycerate (3-PGA). PSAT then catalyses the conversion of PHP to O-phospho L-serine (OPLS), and finally PSP catalyses the cleavage of the phosphate moiety from OPLS to produce L-serine. In many organisms, but not in humans, this L-serine is further converted to cysteine with the help of the enzymes serine acetyltransferase (SAT) (EC 2.3.1.30) and O-acetyl serine sulfhydrylase (OASS) [12]. The absence of this final step from the pathway in humans makes this step an attractive target for a drug that could combat the parasite without producing side effects.

While most of the enzymes of this serine/cysteine biosynthesis pathway have been well characterised in another protozoan parasite, *Entamoeba histolytica*, in our laboratory [13–16] and the effectiveness of inhibitors of this pathway has been reported in some organisms [17], this biosynthesis of phosphorylated serine is not well established in *T. vaginalis*, and its genome appears to be lacking phosphoserine phosphatase (PSP, EC 3.1.3.3) as well as the enzymes for the conversion of L-serine to cysteine [18]. *T. vaginalis* also appears to lack all four

* Corresponding author.

http://dx.doi.org/10.1016/j.bbagen.2016.04.013
0304-4165/

the FEBS Journal

Ligand-induced conformation changes drive ATP hydrolysis and function in SMARCAL1

Meghna Gupta, Mohit Mazumder, Karthik Dhatchinamoorthy[†], Macmillan Nongkhlaw[‡], Dominic Thangminlen Haokip, Samudrala Gourinath, Sneha Sudha Komath and Rohini Muthuswami

School of Life Sciences, Jawaharlal Nehru University, New Delhi, India

Keywords
chromatin remodelling; helicases; Schimke immuno-osseous dysplasia; SMARCAL1; SWI2/SNF2 proteins

Correspondence
R. Muthuswami, Room # 333, School of Life Sciences, JNU, New Delhi 110067, India
Fax: + 91 11 26742558
Tel: + 91 11 26704154
E-mail: rohini_m@mail.jnu.ac.in

Present address
[†]Stowers Institute for Medical Research, Kansas City, MO 64110, USA
[‡]Microbiology Laboratory, North Eastern Hill University, Shillong-793022, Meghalaya, India

Meghna Gupta and Mohit Mazumder contributed equally to this paper

(Received 9 April 2015, revised 3 July 2015, accepted 16 July 2015)

doi:10.1111/febs.13382

Mutations and deletions in SMARCAL1, an SWI2/SNF2 protein, cause Schimke immuno-osseous dysplasia (SIOD). SMARCAL1 preferentially binds to DNA molecules possessing double-stranded to single-stranded transition regions and mediates annealing helicase activity. The protein is critical for alleviating replication stress and maintaining genome integrity. In this study, we have analysed the ATPase activity of three mutations – A468P, I548N and S579L – present in SIOD patients. These mutations are present in RecA-like domain I of the protein. Analysis using active DNA-dependent ATPase A domain (ADAAD), an N-terminal deleted construct of bovine SMARCAL1, showed that all three mutants were unable to hydrolyse ATP. Conformational studies indicated that the α-helix and β-sheet content of the mutant proteins was altered compared to the wild-type protein. Molecular simulation studies confirmed that major structural changes had occurred in the mutant proteins. These changes included alteration of a loop region connecting motif Ia and II. As motif Ia has been implicated in DNA binding, ligand binding studies were done using fluorescence spectroscopy. These studies revealed that the K_d for protein–DNA interaction in the presence of ATP was indeed altered in the case of mutant proteins compared to the wild-type. Finally, *in vivo* studies were done to complement the *in vitro* and *in silico* studies. The results from these experiments demonstrate that mutations in human SMARCAL1 that result in loss in ATPase activity lead to increased replication stress and therefore possibly manifestation of SIOD.

Introduction

SMARCAL1 is a distant member of the SWI2/SNF2 family of ATPases, possessing annealing helicase activity and playing a role in DNA repair by stabilizing the replication fork [1–6]. The protein contains the characteristic helicase motifs essential for DNA binding and ATP hydrolysis [7–9]. Mutations in SMARCAL1 cause Schimke immuno-osseous dysplasia (SIOD), a multi-system disorder characterized by spondyloepiphyseal dysplasia, renal dysfunction and T-cell immunodeficiency [10]. Interestingly most of the mutations in SMARCAL1 leading to SIOD map to the helicase motifs – Q, I, Ia, II, III, IV, V and VI – present in the C-terminus region of SMARCAL1 [10]. Patients with nonsense, frameshift, missense and

Abbreviations
ADAAD, active DNA-dependent ATPase A domain; MD, molecular dynamics; RMSF, root mean square fluctuation; SASA, solvent accessibility surface area; SIOD, Schimke immuno-osseous dysplasia.

Crystal Structure of Calcium Binding Protein-5 from *Entamoeba histolytica* and Its Involvement in Initiation of Phagocytosis of Human Erythrocytes

Sanjeev Kumar[1,2], **Saima Aslam[1]**, **Mohit Mazumder[1]**, **Pradeep Dahiya[3]**, **Aruna Murmu[1]**, **Babu A. Manjasetty[4,5]**, **Rana Zaidi[2]**, **Alok Bhattacharya[1]**, **S. Gourinath[1]***

1 School of Life Sciences, Jawaharlal Nehru University, New Delhi, India, **2** Department of Biochemistry, Jamia Hamdard, New Delhi, India, **3** Plant Mediator Lab, National Institute of Plant & Genome Research, New Delhi, Delhi, India, **4** European Molecular Biology Laboratory, Grenoble Outstation, France, **5** Unit for Virus Host-Cell Interactions, Université Grenoble Alpes - EMBL-CNRS, France

Abstract

Entamoeba histolytica is the etiological agent of human amoebic colitis and liver abscess, and causes a high level of morbidity and mortality worldwide, particularly in developing countries. There are a number of studies that have shown a crucial role for Ca^{2+} and its binding protein in amoebic biology. EhCaBP5 is one of the EF hand calcium-binding proteins of *E. histolytica*. We have determined the crystal structure of EhCaBP5 at 1.9 Å resolution in the Ca^{2+}-bound state, which shows an unconventional mode of Ca^{2+} binding involving coordination to a closed yet canonical EF-hand motif. Structurally, EhCaBP5 is more similar to the essential light chain of myosin than to Calmodulin despite its somewhat greater sequence identity with Calmodulin. This structure-based analysis suggests that EhCaBP5 could be a light chain of myosin. Surface plasmon resonance studies confirmed this hypothesis, and in particular showed that EhCaBP5 interacts with the IQ motif of myosin 1B in calcium independent manner. It also appears from modelling of the EhCaBP5-IQ motif complex that EhCaBP5 undergoes a structural change in order to bind the IQ motif of myosin. This specific interaction was further confirmed by the observation that EhCaBP5 and myosin 1B are colocalized in *E. histolytica* during phagocytic cup formation. Immunoprecipitation of EhCaBP5 from total *E. histolytica* cellular extract also pulls out myosin 1B and this interaction was confirmed to be Ca^{2+} independent. Confocal imaging of *E. histolytica* showed that EhCaBP5 and myosin 1B are part of phagosomes. Overexpression of EhCaBP5 increases slight rate (~20%) of phagosome formation, while suppression reduces the rate drastically (~55%). Taken together, these experiments indicate that EhCaBP5 is likely to be the light chain of myosin 1B. Interestingly, EhCaBP5 is not present in the phagosome after its formation suggesting EhCaBP5 may be playing a regulatory role.

Citation: Kumar S, Aslam S, Mazumder M, Dahiya P, Murmu A, et al. (2014) Crystal Structure of Calcium Binding Protein-5 from *Entamoeba histolytica* and Its Involvement in Initiation of Phagocytosis of Human Erythrocytes. PLoS Pathog 10(12): e1004532. doi:10.1371/journal.ppat.1004532

Editor: William A. Petri, Jr., University of Virginia Health System, United States of America

Received January 13, 2014; **Accepted** October 20, 2014; **Published** December 11, 2014

Funding: Funding is supported by Department of Biotechnology, Government of India. (http://www.dbtindia.gov.in/index.asp) The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: samudralag@yahoo.com

These authors contributed equally to this work.

Introduction

Entamoeba histolytica is the etiological agent of amoebiasis (intestinal as well as extra-intestinal), which results in a high level of morbidity and mortality worldwide, particularly in developing countries [1,2]. A number of studies have shown that Ca^{2+} and its binding proteins are centrally involved in amoebic pathogenesis and that cytolytic activity can be blocked by Ca^{2+} channel blockers or treatment with EGTA [3]. Genomic analysis of *E. histolytica* indicates the presence of 27 genes encoding multiple EF-hand calcium-binding proteins (CaBPs) [4]. The presence of such a large number of CaBPs suggests that this organism has a complex and extensive calcium signalling system [4].

One of the Ca^{2+} sensing proteins of *E. histolytica*, EhCaBP1, has been extensively characterised, both structurally and functionally. EhCaBP1 was found to be involved in cytoskeleton dynamics and is associated with phagocytic cup formation in a Ca^{2+} independent manner [5,6]. The binding of Ca^{2+} to EhCaBP1 is necessary for the transition of phagocytic cups to phagosomes [7]. EhCaBP1 is recruited to phagocytic cups by the novel protein kinase EhC2PK [8]. The crystal structure of EhCaBP1 shows an unusual trimeric arrangement of EF-hand motifs [9]. The structure of the N-terminal lobe of EhCaBP1 displays a similar trimeric organization of EF-hand motifs as observed in the full length molecule. Lowering the pH to below physiological levels was shown to cause a trimer to monomer transition [10]. Moreover, various metal ions have been shown to impart flexibility and plasticity to the EF-hand motifs of EhCaBP1 [11].

We (and others) are systematically investigating the structure-function relationship of other calcium binding proteins of *E. histolytica* as well in order to understand their roles in amoebic biology and pathogenesis. Recently, an NMR structure of the calmodulin-like calcium-binding protein EhCaBP3 has been

EhCoactosin Stabilizes Actin Filaments in the Protist Parasite *Entamoeba histolytica*

Nitesh Kumar[1], Somlata[1], Mohit Mazumder[1], Priyanka Dutta[2], Sankar Maiti[2], Samudrala Gourinath[1]*

1 School of Life Sciences, Jawaharlal Nehru University, New Delhi, India, **2** Indian Institute of Science Education and Research, Kolkata, India

Abstract

Entamoeba histolytica is a protist parasite that is the causative agent of amoebiasis, and is a highly motile organism. The motility is essential for its survival and pathogenesis, and a dynamic actin cytoskeleton is required for this process. EhCoactosin, an actin-binding protein of the ADF/cofilin family, participates in actin dynamics, and here we report our studies of this protein using both structural and functional approaches. The X-ray crystal structure of EhCoactosin resembles that of human coactosin-like protein, with major differences in the distribution of surface charges and the orientation of terminal regions. According to *in vitro* binding assays, full-length EhCoactosin binds both F- and G-actin. Instead of acting to depolymerize or severe F-actin, EhCoactosin directly stabilizes the polymer. When EhCoactosin was visualized in *E. histolytica* cells using either confocal imaging or total internal reflectance microscopy, it was found to colocalize with F-actin at phagocytic cups. Over-expression of this protein stabilized F-actin and inhibited the phagocytic process. EhCoactosin appears to be an unusual type of coactosin involved in *E. histolytica* actin dynamics.

Citation: Kumar N, Somlata, Mazumder M, Dutta P, Maiti S, et al. (2014) EhCoactosin Stabilizes Actin Filaments in the Protist Parasite *Entamoeba histolytica*. PLoS Pathog 10(9): e1004362. doi:10.1371/journal.ppat.1004362

Editor: William A. Petri, Jr., University of Virginia Health System, United States of America

Received February 27, 2014; **Accepted** July 28, 2014; **Published** September 11, 2014

Funding: This work is supported by funding from University Grants Commission (www.ugc.ac.in), INSPIRE Department of Science and Technology (www.inspire-dst.gov.in/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: samudralag@yahoo.com

These authors contributed equally to this work.

Introduction

Human amoebiasis is caused by the protist parasite *E. histolytica*. The parasite is highly motile and displays high level of phagocytic activity in the trophozoite stage. Motility and phagocytosis are essential processes for the survival and invasion of host tissues by the parasite, and largely depends on a highly dynamic actin cytoskeleton. Moreover, there are other processes, such as phagocytosis that also require dynamic actin filament reorganization. Molecular mechanisms that regulate actin dynamics in *E. histolytica* have not been studied in detail. Preliminary investigations suggest an overall similarity with those described in other eukaryotic cells, but with crucial differences. For example, a number of calcium-sensing calcium-binding proteins appear to directly regulate actin recruitment and dynamics [1,2,3]. Several actin-binding proteins are encoded by the *E. histolytica* genome and many of these proteins are homologs of those that have been studied in other systems. Not many of these amebic actin-binding proteins have been characterized. Understanding structural-functional relationship of these proteins would help to decipher mechanisms of actin dynamics in *E. histolytica*.

In *E. histolytica* as well as many other cells, actin dynamics involves both assembly and disassembly of filaments regulated by several actin-binding proteins. The actin-binding protein coactosin was first identified in *Dictyostelium discoidedeum* and has been classified as a member of actin depolymerising factor (ADF)/cofilin family [4]. The ADF/cofilin family members are expressed in all eukaryotes studied to date. The human coactosin-like protein (HCLP) binds F-actin and interferes with capping of filaments. However it does not affect actin polymerisation [5]. HCLP is also known to bind 5-lipooxygenase [6]. The binding of members of the ADF/cofilin family to the F-actin results in severing and depolymerisation of F-actin [7]. However the precise function of this family may vary from actin nucleation to severing depending on the cellular concentration gradient of cofilin [7].

The *E. histolytica* genome contains only one copy of the coactosin gene, whose product we refer to as EhCoactosin. Since the role of EhCoactosin in the actin dynamics of *E. histolytica* has not been previously investigated, we have carried out structural and functional analyses of this protein and present the results here. They show that a single conserved ADF homology domain of EhCoactosin is involved in binding F-actin, and that F-actin is stabilized when EhCoactosin is bound. Moreover, mutation of conserved lysine 75 to alanine does not result in loss of F-actin binding, in contrast to that observed in the case of HCLP, and the binding of this mutant EhCoactosin yields a similar level of F-actin stabilization as does the binding of native EhCoactosin. But deletion of complete F-loop completely abolishes G-actin binding with loss of F-actin stabilization activity, albeit still binds to F-actin. We also propose a mechanism for the binding of EhCoactosin to actin based on a structural model obtained by X-ray crystallography. Overall our results suggest that EhCoactosin displays some features not seen in coactosin from other organisms.

Results

Motility and phagocytosis are important processes for biology of *E. histolytica* as these are involved in providing nutrition and

Journal of General Virology (2014), **95**, 1591–1602 DOI 10.1099/vir.0.064923-0

Mutational analysis of the helicase domain of a replication initiator protein reveals critical roles of Lys 272 of the B′ motif and Lys 289 of the β-hairpin loop in geminivirus replication

Biju George,[1,2] Rajrani Ruhel,[1] Mohit Mazumder,[1] Veerendra Kumar Sharma,[1] Swatantra Kumar Jain,[2] Samudrala Gourinath[1] and Supriya Chakraborty[1]

[1]School of Life Sciences, Jawaharlal Nehru University, New Delhi, India

[2]Department of Biotechnology, Jamia Hamdard University, New Delhi, India

Correspondence
Supriya Chakraborty
supriyachakrasls@yahoo.com

Received 21 February 2014
Accepted 9 April 2014

Replication initiator protein (Rep) is indispensable for rolling-circle replication of geminiviruses, a group of plant-infecting circular ssDNA viruses. However, the mechanism of DNA unwinding by circular ssDNA virus-encoded helicases is unknown. To understand geminivirus Rep function, we compared the sequence and secondary structure of Rep with those of bovine papillomavirus E1 and employed charged residue-to-alanine scanning mutagenesis to generate a set of single-substitution mutants in Walker A (K227), in Walker B (D261, 262), and within or adjacent to the B′ motif (K272, K286 and K289). All mutants were asymptomatic and viral accumulation could not be detected by Southern blotting in both tomato and *N. benthamiana* plants. Furthermore, the K272 and K289 mutants were deficient in DNA binding and unwinding. Biochemical studies and modelling data based on comparisons with the known structures of SF3 helicases suggest that the conserved lysine (K289) located in a predicted β-hairpin loop may interact with ssDNA, while lysine 272 in the B′ motif (K272) located on the outer surface of the protein is presumably involved in coupling ATP-induced conformational changes to DNA binding. To the best of our knowledge, this is the first time that the roles of the B′ motif and the adjacent β-hairpin loop in geminivirus replication have been elucidated.

INTRODUCTION

Geminiviruses cause devastating diseases in a wide range of crop plants worldwide (Stanley, 1983; Fauquet *et al.*, 2003). These viruses, which contain circular ssDNA, are either monopartite or bipartite (i.e. they possess single-component or two-component genomes, respectively). Tomato leaf curl Gujarat virus (ToLCGuV; genus *Begomovirus*, family *Geminiviridae*), is one of the most predominant monopartite begomoviruses causing severe losses to tomato production in the Indian subcontinent (Chakraborty *et al.*, 2003; Chakraborty, 2008).

The geminivirus genome (~2.8 kb) possesses a stem–loop secondary structural element and a direct repeat sequence that function as the origin of replication (ori) and the binding site for replication initiator protein (Rep), respectively. Binding of Rep to the origin leads to strand- and site-specific nicking of viral DNA in an ATP-independent manner. Rep remains covalently linked to the 5′ end of nicked DNA, while the 3′-hydroxyl group is used for the synthesis of the nascent strand (Orozco *et al.*, 1997). Though the detailed three-dimensional structure of Rep is yet to be determined, the protein is known to possess modular functions. The N-terminal region of Rep possesses site-specific nicking, ligation and DNA binding activities (Fontes *et al.*, 1992; Orozco *et al.*, 1997; Chatterji *et al.*, 2000), while the C terminus (aa 120–361) functions autonomously as a 3′ to 5′ helicase (Choudhury *et al.*, 2006; Clérot & Bernardi, 2006). Nonetheless, mechanistic details of Rep-mediated DNA unwinding are currently unknown.

Comparative sequence alignments of the geminivirus Rep proteins have shown that they belong to the SF3 helicase family (Koonin, 1993). Helicases of this family possess three conserved signature motifs: Walker A [involved in ATP binding; GxxxxGK(T/S)], Walker B (involved in ATP hydrolysis; DxxD or xxxxDD), and motif C (a conserved asparagine residue which interacts with the gamma Pi of ATP and an ‘apical’ water molecule). The B′ [$(K/R)x_{3\text{-}4}G_{X7\text{-}8}K$] motif, located between Walker B and motif C, has been identified in SF3 helicases only. The B′ motif has been

One supplementary table and three supplementary figures are available with the online version of this paper.

Biochimica et Biophysica Acta 1830 (2013) 4573–4583

Contents lists available at SciVerse ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbagen

Molecular basis of ligand recognition by OASS from *E. histolytica*: Insights from structural and molecular dynamics simulation studies

Isha Raj, Mohit Mazumder, Samudrala Gourinath *

School of Life Sciences, Jawaharlal Nehru University, New Delhi 110067, India

ARTICLE INFO

Article history:
Received 26 February 2013
Received in revised form 8 May 2013
Accepted 29 May 2013
Available online 6 June 2013

Keywords:
Cysteine biosynthetic pathway
O-acetyl serine sulfhydrylase
Active site cleft
Inhibition
Ligand binding
Conformational change

ABSTRACT

Background: O-acetyl serine sulfhydrylase (OASS) is a pyridoxal phosphate (PLP) dependent enzyme catalyzing the last step of the cysteine biosynthetic pathway. Here we analyze and investigate the factors responsible for recognition and different conformational changes accompanying the binding of various ligands to OASS.
Methods: X ray crystallography was used to determine the structures of OASS from *Entamoeba histolytica* in complex with methionine (substrate analog), isoleucine (inhibitor) and an inhibitory tetra-peptide to 2.00 Å, 2.03 Å and 1.87 Å resolutions, respectively. Molecular dynamics simulations were used to investigate the reasons responsible for the extent of domain movement and cleft closure of the enzyme in presence of different ligands.
Results: Here we report for the first time an OASS-methionine structure with an unmutated catalytic lysine at the active site. This is also the first OASS structure with a closed active site lacking external aldimine formation. The OASS-isoleucine structure shows the active site cleft in open state. Molecular dynamics studies indicate that cofactor PLP, N88 and G192 form a triad of energy contributors to close the active site upon ligand binding and orientation of the Schiff base forming nitrogen of the ligand is critical for this interaction.
Conclusions: Methionine proves to be a better binder to OASS than isoleucine. The β branching of isoleucine does not allow it to reorient itself in suitable conformation near PLP to cause active site closure.
General significance: Our findings have important implications in designing better inhibitors against OASS across all pathogenic microbial species.

1. Introduction

Growth and survival of the protozoan parasite *Entamoeba histolytica* are critically dependent upon the cysteine biosynthetic pathway. Cysteine, which is the product of this pathway, is the only anti-oxidative thiol in *Entamoeba histolytica* and plays an important role in maintaining the redox balance in this organism [1]. This amino acid is important for optimal growth of *Entamoeba* and is essential for its attachment and survival under oxidative stress [2–6].

The de novo cysteine biosynthetic pathway starts with serine acetyl transferase (SAT, EC 2.3.1.30) catalyzing the formation of O-acetyl serine (OAS) from acetyl Co-A and serine. OAS is then converted to cysteine by the addition of sulfide and elimination of acetate in a reaction catalyzed by O-acetyl serine sulfhydrylase (OASS, EC 2.5.1.47). OASS follows a ping pong kinetic mechanism where the conserved catalytic lysine residue forms an internal aldimine with PLP in the native state. OAS substitutes for lysine at the active site and forms an external Schiff base with PLP, followed by β elimination in which acetate is released and a proton is abstracted from the α position [7]. This leads to the formation of the α amino acrylate intermediate covalently linked to PLP (Fig. 1). Nucleophilic attack of the second substrate, sulfide, on the β carbon of the amino acrylate intermediate re-protonates the α carbon, resulting in cysteine bound as an external Schiff base. The product is then released, restoring the internal aldimine.

The structure of OASS from *E. histolytica* (EhOASS) in its native (i.e unmutated/unliganded) form, as well as with cysteine bound, has been reported [8]. The conformation of EhOASS belongs to the type II fold of PLP dependent enzymes [9,10], similar to that in other plant and bacterial OASSs [11–14]. Structural and biochemical studies have shown that in addition to the substrate OAS, OASS can bind to cysteine, its product, and to methionine, a substrate analog [8,14,15]. OASS activity is regulated both by its metabolites and by interaction with SAT, the other enzyme of the cysteine biosynthetic pathway. The SAT C-terminal peptide has been shown to interact with the OASS active site to inhibit its activity [16–19]. The common feature of all these SATs is the presence of a conserved Ile at its C-terminal end. Cysteine synthase complex does not form in *E. histolytica*, despite the presence of isoleucine at the C-terminal end of SAT1 in this organism [20].

Earlier studies to determine the conformational changes taking place upon substrate binding have mostly employed methionine as a substrate analog and a modified OASS, where the catalytic Lys was mutated to Ala. The structure of an OASS K41A mutant from *Salmonella typhimurium* in complex with methionine revealed methionine bound in an external aldimine (EA) linkage with PLP and accompanied with

* Corresponding author. Tel.: +91 11 26704513; fax: +91 11 26187338.
E-mail address: sgourinath@mail.jnu.ac.in (S. Gourinath).

0304-4165/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.bbagen.2013.05.041

JB Journals.ASM.org

Crystal Structure and Mode of Helicase Binding of the C-Terminal Domain of Primase from *Helicobacter pylori*

Syed Arif Abdul Rehman,[a] Vijay Verma,[b] Mohit Mazumder,[a] Suman K. Dhar,[b] S. Gourinath[a]

School of Life Sciences, Jawaharlal Nehru University, New Delhi, India[a]; Special Center for Molecular Medicine, Jawaharlal Nehru University, New Delhi, India[b]

To better understand the poor conservation of the helicase binding domain of primases (DnaGs) among the eubacteria, we determined the crystal structure of the *Helicobacter pylori* DnaG C-terminal domain (*Hp*DnaG-CTD) at 1.78 Å. The structure has a globular subdomain connected to a helical hairpin. Structural comparison has revealed that globular subdomains, despite the variation in number of helices, have broadly similar arrangements across the species, whereas helical hairpins show different orientations. Further, to study the helicase-primase interaction in *H. pylori*, a complex was modeled using the *Hp*DnaG-CTD and *Hp*DnaB-NTD (helicase) crystal structures using the *Bacillus stearothermophilus Bst*DnaB-*Bst*DnaG-CTD (helicase-primase) complex structure as a template. By using this model, a nonconserved critical residue Phe534 on helicase binding interface of DnaG-CTD was identified. Mutation guided by molecular dynamics, biophysical, and biochemical studies validated our model. We further concluded that species-specific helicase-primase interactions are influenced by electrostatic surface potentials apart from the critical hydrophobic surface residues.

Replication of chromosomal DNA is generally a universal process that requires a high degree of accuracy and precision to maintain fidelity in the transmission of genetic material from one generation to the next (1–4). This unique process involves multiprotein complexes that help to check the inevitable errors associated with DNA replication (5–7). Interference with any of these protein-DNA and protein-protein interactions may lead to numerous problems, including unviable offspring. Eubacterial DnaG primase is a single-stranded DNA (ssDNA)-dependent RNA polymerase responsible for the synthesis of oligonucleotide primers needed for DNA replication (8). Primase is recruited once or twice on the leading strand, in contrast to the lagging strand, where it is recruited several times (9, 10). DnaG primase also plays an important role in tuning the synthesis (11, 12). The eubacterial DnaG primase has three domains. The N-terminal domain (NTD) is involved in template DNA recognition and contains a zinc binding domain, and the central catalytic domain synthesizes oligonucleotide primers. The only known function of the C-terminal domain (CTD), also known as the helicase binding domain (HBD), is to interact with helicase at the replication fork. Of these domains, the CTD is least conserved (13). The HBD/CTD is sufficient to bind and stimulate the activities of DnaB helicase (3, 14). The stability of the interaction between DnaG primase and DnaB helicase varies among eubacteria. In *Escherichia coli* the interaction has been reported to be weak (9, 15, 16), whereas in *Bacillus stearothermophilus* the interaction is so strong that it can be purified on a gel filtration column (17). We have recently reported a moderate level of interaction between these proteins in *Helicobacter pylori* (14). The full-length structure of DnaG primase has yet to be determined, although the structures of individual domains have been reported. Primase C-terminal domain crystal and solution structures are known from *E. coli* (10) and *B. stearothermophilus* (13). The crystal structures of the zinc binding domain alone and together with RNA polymerase domain structure were determined in *B. stearothermophilus* (18) and *Aquifex aeolicus* (19), respectively. RNA polymerase domain and its complex with ssDNA crystal structures for *E. coli* were recently published (20, 21). Recently, a medium resolution helicase-primase complex structure was reported, and this study provided insight into the interaction pattern of the DnaG CTD with DnaB helicase (22). Since the helicase-binding domains in primases are poorly conserved, high-resolution structures from different organisms will be helpful for understanding the mechanism of interaction between DnaG and DnaB. *H. pylori* infection is present in half of the world's human population (23). It causes diverse diseases of the stomach, from chronic gastritis to mucosa-associated lymphoid tissue lymphomas (24, 25). Interestingly, recent work has shown that this organism may also have a role in autoimmune thrombocytopenia (26), Guillain-Barre syndrome (27), and in Alzheimer's disease (28) and in strokes (29). These diseases and especially ailments, such as persistent diarrhea, peptic ulcer, and gastric cancer, may be mitigated if a way can be found to eradicate *H. pylori*. Since the present therapeutic approach targeting this organism is not very effective, and the conditions in developing countries are suitable for this organism to flourish, a better therapeutic approach is urgently needed. Since the initiation of replication is a crucial step in reproduction of an organism, the structural and functional studies of this target process is important for future drug development.

The proteins involved in DNA replication and repair in *H. pylori* have been reviewed recently (30). Several DNA replication proteins, such as the initiator proteins DnaA and Hob-A (31–34), the replicative helicase DnaB and its unique dodecameric architecture (35–38), and the single-stranded DNA-binding protein

Received 21 January 2013 Accepted 6 April 2013

Published ahead of print 12 April 2013

Address correspondence to S. Gourinath, sgourinath@mail.jnu.ac.in, or Suman K. Dhar, skdhar2002@yahoo.co.in.

S.A.A.R. and V.V. are equally contributing first authors.

Supplemental material for this article may be found at http://dx.doi.org/10.1128/JB.00091-13.

doi:10.1128/JB.00091-13

Single Residue Mutation in Active Site of Serine Acetyltransferase Isoform 3 from *Entamoeba histolytica* Assists in Partial Regaining of Feedback Inhibition by Cysteine

Sudhir Kumar, Mohit Mazumder, Sudhaker Dharavath, S. Gourinath*

School of Life Sciences, Jawaharlal Nehru University, New Delhi, Delhi, India

Abstract

The cysteine biosynthetic pathway is essential for survival of the protist pathogen *Entamoeba histolytica*, and functions by producing cysteine for countering oxidative attack during infection in human hosts. Serine acetyltransferase (SAT) and O-acetylserine sulfhydrylase (OASS) are involved in cysteine biosynthesis and are present in three isoforms each. While EhSAT1 and EhSAT2 are feedback inhibited by end product cysteine, EhSAT3 is nearly insensitive to such inhibition. The active site residues of EhSAT1 and of EhSAT3 are identical except for position 208, which is a histidine residue in EhSAT1 and a serine residue in EhSAT3. A combination of comparative modeling, multiple molecular dynamics simulations and free energy calculation studies showed a difference in binding energies of native EhSAT3 and of a S208H-EhSAT3 mutant for cysteine. Mutants have also been generated *in vitro*, replacing serine with histidine at position 208 in EhSAT3 and replacing histidine 208 with serine in EhSAT1. These mutants showed decreased affinity for substrate serine, as indicated by K_m, compared to the native enzymes. Inhibition kinetics in the presence of physiological concentrations of serine show that IC50 of EhSAT1 increases by about 18 folds from 9.59 μM for native to 169.88 μM for H208S-EhSAT1 mutant. Similar measurements with EhSAT3 confirm it to be insensitive to cysteine inhibition while its mutant (S208H-EhSAT3) shows a gain of cysteine inhibition by 36% and the IC50 of 3.5 mM. Histidine 208 appears to be one of the important residues that distinguish the serine substrate from the cysteine inhibitor.

Citation: Kumar S, Mazumder M, Dharavath S, Gourinath S (2013) Single Residue Mutation in Active Site of Serine Acetyltransferase Isoform 3 from *Entamoeba histolytica* Assists in Partial Regaining of Feedback Inhibition by Cysteine. PLoS ONE 8(2): e55932. doi:10.1371/journal.pone.0055932

Editor: Rajagopal Subramanyam, University of Hyderabad, India

Received October 19, 2012; **Accepted** January 3, 2013; **Published** February 21, 2013

Funding: SK and MM thank University Grants Commission (UGC) and DBT for fellowship respectively. The authors thank Programme support on Molecular Parasitology by Department of Biotechnology, and Council of Scientific and Industrial research, Government of India. The authors thank UGC resource networking and Department of Science and Technology-Fund for Improvement of S&T Infrastructure in Higher Educational Institutions funding for departmental central facility. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: sgourinath@mail.jnu.ac.in

Introduction

Serine acetyltransferase (SAT) (EC 2.3.1.30) which is the first member of the two-step cysteine biosynthetic pathway, catalyzes the formation of O-acetylserine (OAS) by transferring the acetyl group of acetyl Coenzyme A to serine (Ser) [1]. The SAT structure includes, in its C-terminal domain, a well conserved pair of so-called left handed parallel β-sheet helices (LβH), which arise due to a repeat sequence of [LIV]-[GAED]-X2-[STAV]-X [2] and also contribute to the formation of the active site. Comparison of the SAT structures available in native as well substrate/inhibitor bound forms shows that the residues involved in the substrate binding are highly conserved. SAT in bacteria and plants combines with the second member of the cysteine biosynthetic pathway, O-Acetyl Serine Sulfhydrylase (OASS) to form a cysteine synthase (CS) complex [3]. CS complex formation, which is favored when sufficient sulfur is available, is a part of the regulatory mechanism of the pathway where the activity of SAT increases and that of OASS decreases. Decrease in sulfide levels and excess production of OAS result in dissociation of the CS complex and an increase in OASS activity. Another level of regulation results from feedback inhibition of SAT by the cysteine (Cys) end product. In all of the organisms, where this pathway has been explored, most of the SAT isoforms are known to be competitively inhibited by cysteine, while a few SAT isoforms were also reported to exhibit a loss of inhibition by cysteine [3,4,5]. CS complex formation is absent in *E. histolytica*, and the feedback inhibition seems to be the only regulatory pathway in the protist pathogen. *E. histolytica* thus appears to be solely dependent on cysteine for anti-oxidative defense [6,7,8].

There are three isoforms of SAT in *Entamoeba histolytica*. EhSAT2 and EhSAT3 share 73% and 48% sequence identity respectively with EhSAT1 [4]. EhSAT1 was first characterized in *Entamoeba* by Nozaki and colleagues and they proposed the loss of interaction between SAT and OASS [8]. Hussain and colleagues characterized the remaining two isoforms and showed that feedback inhibition by Cys is different for all the three EhSAT isoforms. The EhSAT1 and EhSAT2 isoforms were inhibited by about 95% and 75% respectively, but EhSAT3 remained insensitive to cysteine even at high concentrations and in the presence of physiological concentrations of serine (3 mM) [4]. The crystal structure of EhSAT1 reported by our group, established the

The GPI Anchor Signal Sequence Dictates the Folding and Functionality of the Als5 Adhesin from *Candida albicans*

Mohammad Faiz Ahmad[1,2], Bhawna Yadav[1], Pravin Kumar[3], Amrita Puri[4], Mohit Mazumder[1], Anwar Ali[2], Samudrala Gourinath[1], Rohini Muthuswami[1], Sneha Sudha Komath[1]*

1 School of Life Sciences, Jawaharlal Nehru University, New Delhi, India, **2** Department of Chemistry, Jamia Millia Islamia, New Delhi, India, **3** Department of Plant Molecular Biology, University of Delhi, New Delhi, India, **4** Invisible Sentinel, Biotech Company, Philadelphia, Pennsylvania, United States of America

Abstract

Background: Proteins destined to be Glycosylphosphatidylinositol (GPI) anchored are translocated into the ER lumen completely before the C-terminal GPI anchor attachment signal sequence (SS) is removed by the GPI-transamidase and replaced by a pre-formed GPI anchor precursor. Does the SS have a role in dictating the conformation and function of the protein as well?

Methodology/Principal Findings: We generated two variants of the Als5 protein without and with the SS in order to address the above question. Using a combination of biochemical and biophysical techniques, we show that in the case of Als5, an adhesin of *C. albicans*, the C-terminal deletion of 20 amino acids (SS) results in a significant alteration in conformation and function of the mature protein.

Conclusions/Significance: We propose that the locking of the conformation of the precursor protein in an alternate conformation from that of the mature protein is one probable strategy employed by the cell to control the behaviour and function of proteins intended to be GPI anchored during their transit through the ER.

Citation: Ahmad MF, Yadav B, Kumar P, Puri A, Mazumder M, et al. (2012) The GPI Anchor Signal Sequence Dictates the Folding and Functionality of the Als5 Adhesin from *Candida albicans*. PLoS ONE 7(4): e35305. doi:10.1371/journal.pone.0035305

Editor: Scott G. Filler, David Geffen School of Medicine at University of California Los Angeles, United States of America

Received December 9, 2011; **Accepted** March 13, 2012; **Published** April 11, 2012

Funding: This work was supported in part by grants to SSK from DBT-India (grant number: BT/PR10689/10/616/2008) (URL:http://dbtindia.nic.in), UGC-RNW (grant number: 34–282/2008(SR)) (URL:http://www.ugc.ac.in/) and the SLS LRE. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. No other external funding was received for this study.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: sskomath@yahoo.com

Introduction

A wide variety of proteins are known to be anchored to the extra-cytoplasmic leaflet of the plasma membrane by glycosylphosphatidylinositol (GPI) anchors and defects in GPI anchor attachment can have severe consequences for the eukaryotic cell [1]. Proteins destined to be GPI anchored possess a C-terminal signal sequence specific for this modification [2]. Unlike integral membrane proteins that have their transmembrane domains co-translationally inserted into the membrane via the translocon pore, proteins meant to be GPI anchored are completely translocated into the ER lumen [3]. Shortly thereafter, these are acted upon by the GPI-transamidase and have their C-terminal GPI anchor attachment signal sequence (SS) replaced by a pre-formed GPI anchor.

Is the role of the SS confined to being a signal for GPI anchor attachment or does it also control the conformation and function of a protein destined to be GPI anchored? In order to address this question, we chose to study Als5, an adhesin from *Candida albicans*.

ALS5 belongs to the agglutinin-like sequence (*ALS*) family of genes which code for eight adhesins in *Candida albicans*. These adhesins are important for establishment of commensal colonies of the organism in the host as well as in its pathogenesis and virulence under appropriate conditions [4]. Since they are tethered to the membrane via GPI anchors, any defects in GPI anchor biosynthesis can drastically affect the pathogenesis and virulence of the organism [5–7]. Indeed, complete GPI anchors have been shown to be important for morphogenesis, virulence and macrophage-resistance of the organism [7].

Like other members of the Als family of adhesins, Als5 has an N-terminal secretion signal followed by a large immunoglobulin-like domain, a highly conserved Thr-rich segment, a central domain containing variable numbers of tandem repeats of Ser/Thr sequences, a C-terminal Ser/Thr rich stalk and the C-terminal signal sequence for GPI anchor attachment [8]. When heterologously expressed in *S. cerevisiae*, Als5 can make the host cells adhere to basal lamina proteins such as collagen type IV and fibronectin [9]. The protein has also been shown to be capable of mediating endothelial cell invasion and its N-terminal domain has been shown to be important for adherence [10,11]. The protein has a tendency to aggregate and form amyloid-like fibrils; a potential amyloidogenic domain has also been identified [11–13].

In this study, we show that it is possible to express Als5 as a GST-fusion protein in bacterial cells and to purify it using affinity

Multiple Sequence Alignment Based Upon Statistical Approach of Curve Fitting

Vineet Jha, Mohit Mazumder*, Hrishikesh Bhuyan, Ashwani Jha, and Abhinav Nagar

InSilico Biosolution, 103B, North Guwahati College Road, Abhoypur, Near IIT-Guwahati, P.O – College Nagar, North Guwahati – 781031, Assam, India
mazumder.mohit@gmail.com

Abstract. The main objective of our work is to align multiple sequences together on the basis of statistical approach in lieu of heuristics approach. Here we are proposing a novel idea for aligning multiple sequences in which we will be considering the DNA sequences as lines not as strings where each character represents a point in the line. DNA sequences are aligned in such a way that maximum overlap can occur between them, so that we get maximum matching of characters which will be treated as our seeds of the alignment. The proposed algorithm will first find the seeds in the aligning sequences and then it will grow the alignment on the basis of statistical approach of curve fitting using standard deviation.

Keywords: Multiple Sequence Alignment, Sequence Alignment, Word Method, Statistically Optimized Algorithm, Comparative Genome Analysis, Cross Referencing, Evolutionary Relationship.

1 Introduction

Multiple sequence alignment is a crucial prerequisite for biological sequence data analysis.

It is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. A large number of multi-alignment programs have been developed during last twenty years. There are three main considerations in choosing a program: biological accuracy, execution time and memory usage. Biological accuracy is generally the most important concern amongst all. Some of the prominent and accurate programs according to most benchmarks are *CLUSTAL W* [1], *DIALIGN* [2], *T-COFFEE* [3], MAFFT, MUSCLE, PROBCONS . An overview about these tools and other established methods are given [4].

T-COFFEE is a prototypical consistency- based method which is still considered as one of the most accurate program available. MAFFT and MUSCLE have a similar design, building on work done by Gotoh in the 1990s that culminated in the PRRN

* Corresponding author.

S. Chaudhury et al. (Eds.): PReMI 2009, LNCS 5909, pp. 184–192, 2009.

2009 IEEE International Advance Computing Conference (IACC 2009)
Patiala, India, 6–7 March 2009

Upstream Sequence Finder- Tool to Find Out Upstream Element in Various Database or Genome.

Vineet Jha, Mohit Mazumder and Susanta Roy*

***Corresponding address**: Bio Explore, C-5, Hazipark, Dimapur, Nagaland – 797112, India*
***Registered Office**: Bio Explore, G/7, Thongpok Market, Nagabazar, Dimapur, Nagaland – 797112, India.*
Email: susanta.roy@bioexplore.co.in Website: www.bioexplore.co.in

Abstract: Upstream elements are very significant in disclosing the property of the sequence not only they set a signal for the various protein to bind there but also help in locating hidden sequences and their property like TATA box. The whole idea about developing this algorithm is that to find out upstream sequences which carry hidden property like road signs which can alert drivers. In this case protein help user to predict and analyse the upstream sequences. We downloaded the DATABASE file (nucleotide file), query file and did the nBLAST. Then we parse the blast output to filter out full length sequences (sequences which are not truncated either from 5' or 3' end for more than 11 bases).The time complexity of algorithm was improved from exponential time complex to linear time complex by using the divide and conquer approach, where the large database file is divided into smaller files. This algorithm gives good hits and filters out the upstream element. One can even fix the option of having a gap or un-gapped alignment in the database.

1. Introduction

Upstream element: Transcription runs 5' (start) to 3' (end). Anything that sits prior to the 5' of the start site is outside the transcribed region called "upstream" of the start site.

To know the correct functioning of multi cellular organisms one should have the knowledge of complex orchestration of gene regulatory events, which make sure that genes are expressed at the right time, place and level. To a great extent of this regulation occurs at the level of gene transcription, and is mediated by specific interactions between transcription factors and cis-regulatory DNA motifs. Upstream to the transcription start site (TSS), the gene promoter region show regulatory motifs concentration. (For a recent review, see **[1]**).

In 1979, it was discovered that DNA sequences thousands of nucleotides away from a eukaryotic promoter, could trigger transcription from the promoter. Such enhancer sequences hand round as specific binding sites for gene regulatory proteins that enhance transcription. **[3]**

The upstream (towards the 5' region) of a gene is the regulatory region of DNA, providing a control point for regulated gene transcription called as the promoter which contains specific DNA sequences that are recognized by proteins known as transcription factors. These factors bind to the promoter sequences, recruiting RNA polymerase, the enzyme that synthesizes the RNA from the coding region of the gene. [3]

2. Promoter Elements

2.1 Core Promoter - The minimal portion of the promoter required to properly initiate transcription [2, 3].

1. Transcription Start Site (TSS).
2. Approximately -34 (Base pairs).
3. A binding site for RNA polymerase.
4. General transcription factor binding sites.

2.2 Proximal Promoter - The proximal sequence upstream of the gene that tends to contain primary regulatory elements [2]

1. Approximately -250 (Base pairs).
2. Specific transcription factor binding sites

2.3 Prokaryotic promoters - In prokaryotes, the promoter consists of two short sequences at -10 and -35 positions upstream from the transcription start site. [2]

1. The sequence at **-10** is called the Pribnow box [9], or the -10 element, and usually consists of the six nucleotides **TATAAT.** The Pribnow

Cloning, Expression and Functional Characterization of Als5: An Adhesin from *Candida albicans*

Sneha Sudha Komath, Mohammad Faiz Ahmad and Mohit Mazumder
School of Life Sciences
Jawaharlal Nehru University, India

Chapter 21
Structural Biology of Cysteine Biosynthetic Pathway Enzymes

Isha Raj, Sudhir Kumar, Mohit Mazumder, and S. Gourinath

Abstract The cysteine biosynthetic pathway is of central importance for the growth, survival, and pathogenicity of the anaerobic protozoan parasite *Entamoeba histolytica*. This pathway is present across all species but is absent in mammals. Cysteine, the product of this pathway, is the only antioxidative thiol responsible for fighting oxidative stress in *E. histolytica*. Serine acetyl transferase (SAT) and *O*-acetyl serine sulfhydrylase (OASS) are the two enzymes catalyzing the de novo cysteine biosynthetic pathway. In all organisms in which so far this pathway is known to exist, both these enzymes associate to form a regulatory complex, but in *E. histolytica* this complex is not formed. The cysteine biosynthetic pathway has been optimized in this organism to adapt to and fulfill its cysteine requirements. Here we describe recent studies of the structure, function, and complex formation of cysteine biosynthetic enzymes in *E. histolytica*. The findings reveal subtle modifications that lend both cysteine biosynthetic enzymes their unique characteristics to escape inhibitory regulation; allowing *E. histolytica* to maintain high levels of cysteine at all times.

21.1 Cysteine Biosynthetic Pathway: An Overview

The de novo cysteine biosynthetic pathway is of primary importance in anaerobic microorganisms as it incorporates inorganic sulfur into an organic skeleton to produce cysteine. Cysteine serves important roles both as an antioxidative agent and as a source of sulfur for biomolecules such as thiamine, Fe–S clusters, biotin, Co-A, methionine, and various antioxidative thiols (glutathione, mycothiol, trypanothione) [1, 2]. In *Entamoeba histolytica*, the antioxidative role of cysteine is critical as it is the sole thiol responsible for maintaining the redox state in this catalase- and peroxidase-deficient parasitic protozoan [3, 4]. Cysteine deprivation has far-reaching effects in *E. histolytica*. Gene expression analysis has shown that it alters the

I. Raj • S. Kumar • M. Mazumder • S. Gourinath (✉)
Jawaharlal Nehru University, New Delhi, India
e-mail: sgn9@hotmail.com

T. Nozaki, A. Bhattacharya (eds.), *Amebiasis*,
DOI 10.1007/978-4-431-55200-0_21

Turnitin Originality Report

combo by Chap3 Chap1

From Revision 1 (mohit thesis)

Processed on 25-Jul-2017 12:34 IST
ID: 833005204
Word Count: 10871

Similarity Index	Similarity by Source	
11%	Internet Sources:	4%
	Publications:	9%
	Student Papers:	1%

sources:

1 1% match (Internet from 21-Nov-2014)
http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0096202

2 1% match (Internet from 22-Jun-2014)
http://www.coursehero.com/file/2818314/zhouyubin200712phd/

3 1% match (student papers from 29-Mar-2017)
Submitted to University of Hong Kong on 2017-03-29

4 1% match (publications)
Kumar, Sanjeev, Rana Zaidi, and Samudrala Gourinath. "Cloning, purification, crystallization and preliminary crystallographic study of calcium-binding protein 5 from *Entamoeba histolytica*", Acta Crystallographica Section F Structural Biology and Crystallization Communications, 2012.

5 1% match (publications)
Ruchi Jain. "N- and C-Terminal Domains of the Calcium Binding Protein EhCaBP1 of the Parasite Entamoeba histolytica Display Distinct Functions", PLoS ONE, 04/22/2009

6 1% match (publications)
Casadei, Federica <1978>(Cremonini, Dott. Mauro Andrea). "Application of nuclear magnetic resonance spectroscopy for the evaluation of folding variability in calcium binding proteins and its implications in food allergies", Alma Mater Studiorum - Università di Bologna, 2011.

7 1% match (publications)
Permyakov. "Calcium, Calcium-Binding Proteins, and their Major Families", Metalloproteomics, 04/07/2009

< 1% match (publications)